UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

CAIO DOS SANTOS NASCIMENTO

MICROESTADOS CEREBRAIS NO DIAGNÓSTICO DE DISTÚRBIOS MENTAIS: UMA ABORDAGEM MULTIVARIADA DE CLUSTERIZAÇÃO E CLASSIFICAÇÃO DE ALGORITMOS

FORTALEZA

2024

CAIO DOS SANTOS NASCIMENTO

MICROESTADOS CEREBRAIS NO DIAGNÓSTICO DE DISTÚRBIOS MENTAIS: UMA ABORDAGEM MULTIVARIADA DE CLUSTERIZAÇÃO E CLASSIFICAÇÃO DE ALGORITMOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Victor Hugo Costa de Albuquerque

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

N194m Nascimento, Caio dos Santos.
Microestados cerebrais no diagnóstico de distúrbios mentais : uma abordagem multivariada de clusterização e classificação de algoritmos / Caio dos Santos Nascimento. – 2024.
101 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia de Computação, Fortaleza, 2024.
Orientação: Prof. Dr. Victor Hugo Costa de Albuquerque.

1. EEG. 2. Microestados. 3. Clusterização. 4. Classificação. I. Título.

CDD 621.39

CAIO DOS SANTOS NASCIMENTO

MICROESTADOS CEREBRAIS NO DIAGNÓSTICO DE DISTÚRBIOS MENTAIS: UMA ABORDAGEM MULTIVARIADA DE CLUSTERIZAÇÃO E CLASSIFICAÇÃO DE ALGORITMOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Aprovada em:

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br
VICTOR HUGO COSTA DE ALBUQUERQUE
Data: 20/09/2024 08:55:45-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Victor Hugo Costa de Albuquerque (Orientador)
Universidade Federal do Ceará (UFC)

Documento assinado digitalmente
gov.br
RENE RIPARDO CALIXTO
Data: 19/09/2024 18:18:09-0300
Verifique em https://validar.iti.gov.br

Prof. MSc. Renê Ripardo Calixto
Universidade Federal do Ceará (UFC)

Documento assinado digitalmente
gov.br
RANIERE ROCHA GUIMARAES
Data: 20/09/2024 00:55:05-0300
Verifique em https://validar.iti.gov.br

Prof. MSc. Raniere Rocha Guimarães
Instituto Federal do Pará (IFPA)

À minha família, por seu constante investimento em mim. À minha namorada por me dar apoio e incentivo. Com a contribuição de vocês, pude perserverar mesmo em momentos de desmotivação e conseguir concluir toda a graduação.

AGRADECIMENTOS

Ao Prof. Dr. Victor Hugo Costa de Albuquerque pela orientação na produção de artigos e do meu trabalho de conclusão de curso.

Aos meus demais professores que se empenharam em passar o máximo de conhecimento em suas respectivas áreas.

Aos meus amigos que tornaram todo o processo mais leve e divertido, mas também compartilharam conhecimento e contribuíram para o meu desenvolvimento acadêmico e profissional.

À minha namorada que me ajudou a persistir frente às dificuldades enfrentadas nessas últimas etapas.

Aos meus pais que sempre apoiaram no meu trajeto educacional desde a base até o atual ponto do Ensino Superior.

"A inteligência consiste não só no conhecimento, mas também na habilidade de aplicar o conhecimento na prática."

(Aristóteles)

RESUMO

Os transtornos mentais estão sendo diagnosticados em proporções alarmantes em todo o mundo, assumindo características epidêmicas que demonstram a urgência de desenvolver métodos eficazes de prevenção e intervenção. Nesse contexto, o eletroencefalograma (EEG) se destaca como uma das ferramentas mais utilizadas para analisar a atividade elétrica cerebral, permitindo o estudo dos microestados cerebrais. Estes microestados, são considerados biomarcadores promissores, mostrado-se úteis no diagnóstico e no acompanhamento da evolução de distúrbios mentais, contribuindo significativamente para o avanço da literatura médica sobre essas doenças. Neste trabalho, foi realizada uma análise comparativa de diversos algoritmos de clusterização e classificação, integrados em um fluxo de validação de performance, com o objetivo de oferecer uma análise multivariada das métricas em função de diferentes números de microestados. Os resultados da etapa de classificação indicaram que os algoritmos *Atomize and Agglomerative Hierarchical Clustering* (AAHC) e o *Modified K-Means* se destacaram, atingindo métricas em torno de 80% de acurácia. Com base nesses resultados, os números de microestados que apresentaram melhor performance foram selecionados para uma avaliação detalhada da clusterização, utilizando as métricas *Silhouette*, *Calinski-Harabasz*, *Davies-Bouldin* e *Dunn Index*. As análises convergiram para os microestados 2 e 5 como os que proporcionaram os melhores resultados, sugerindo que esses números são os mais adequados para futuras investigações e aplicações clínicas.

Palavras-chave: EEG. Microestados. Clusterização. Classificação.

ABSTRACT

Mental disorders are being diagnosed at alarming rates worldwide, taking on epidemic proportions that highlight the urgent need to develop effective methods of prevention and intervention. In this context, the electroencephalogram (EEG) stands out as one of the most widely used tools for analyzing brain electrical activity, enabling the study of brain microstates. These microstates, considered promising biomarkers, have proven useful in the diagnosis and monitoring of mental disorder progression, significantly contributing to the advancement of medical literature on these diseases. In this study, a comparative analysis of various clustering and classification algorithms was conducted, integrated into a performance validation flow, with the aim of providing a multivariate analysis of metrics in relation to different numbers of microstates. The results of the classification stage indicated that the Atomize and Agglomerative Hierarchical Clustering (AAHC) and the Modified K-Means algorithms stood out, reaching accuracy metrics around 80%. Based on these results, the numbers of microstates that presented the best performance were selected for a detailed evaluation of the clustering, using the Silhouette, Calinski-Harabasz, Davies-Bouldin, and Dunn Index metrics. The analyses converged on microstates 2 and 5 as those that provided the best outcomes, suggesting that these numbers are the most suitable for future investigations and clinical applications.

Keywords: EEG. Microstates. Clustering. Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANNs	Redes Neurais Artificiais
ATP	Adenosina Tri-Fosfato
CNNs	Redes Neurais Convolucionais
DBSCAN	Clusterização espacial baseada em densidade de dados com ruído
EDF	*European Data Format*
EEG	Eletroencefalograma
K-Means	K-Médias
OMS	Organização Mundial da Saúde
OPAS	Organização Pan-Americana de Saúde
OPF	Floresta de Caminhos Ótimos
REM	*Rapid Eye Movement*
SVMs	Máquinas de Vetores de Suporte
TUAB	*Temple University Hospital Abnormal EEG Corpus*
TUEG	*TUH EEG Corpus*

SUMÁRIO

1 INTRODUÇÃO

Essa introdução irá abranger as informações importantes para compreender o contexto em que se aplicará a utilidade desse trabalho, assim como os objetivos motivadores. Serão explicados com mais detalhes os aspectos fisiológicos do cérebro, as dificuldades em executar o exame e interpretar o laudo da eletroencefalografia e o atual uso de algoritmos de aprendizagem de máquina para extrair informações do resultado.

1.1 Contextualização da problemática

Um estudo publicado pela revista científica *Lancet Neurology* mostrou que três bilhões, ou seja, uma em cada três pessoas são afetadas por doenças neurológicas, que lideram as causas de problemas de saúde e incapacidade. Nesse estudo, destaca-se também o aumento em 18% da quantidade total de incapacidade, doença e morte prematura provocadas por condições neurológicas em todo o mundo. (LANCET, 2024)

A Organização Mundial da Saúde (OMS) também publicou sua análise a qual revela que aproximadamente 70% dos casos de doenças neurológicas ocorrem em países de baixa e média renda, sendo responsáveis por cerca de nove milhões de mortes por ano. E ainda, as doenças que mais contribuem para os anos de vida ajustados por incapacidade, referido como *DALYs*, foram o acidente vascular cerebral (42,2%), a enxaqueca (16,3%), a demência (10,4%), a meningite (7,9%) e a epilepsia (5%). Com o intuito de ajudar a melhorar esse cenário, membros da OMS criaram um plano de ação para direcionar os esforços em prol de aperfeiçoar o acesso aos tratamentos e cuidados necessários, a fim de impactar a qualidade de vida das pessoas acometidas por essas doenças. (OMS, 2022)

Voltada para um âmbito mais restrito, a Organização Pan-Americana de Saúde (OPAS) realizou uma pesquisa, considerando apenas os países das Américas, em que atribui o crescimento de mortes e de incapacitação causadas pelos problemas neurológicos ao fato de que a população mundial está vivendo por mais tempo. Dos resultados encontrados, comparando homens e mulheres no ano de 2019, de aproximadamente 533 mil mortes, 60% foram de mulheres e, padronizando os dados para um conjunto de 100 mil habitantes, as taxas de morte variam de alto (nos Estados Unidos com 47.4 mortes por 100 mil habitantes) a baixo (na Venezuela com 6.6 mortes por cem mil habitantes).

Já a análise de anos de vida perdidos devido à morte prematura (referido no estudo

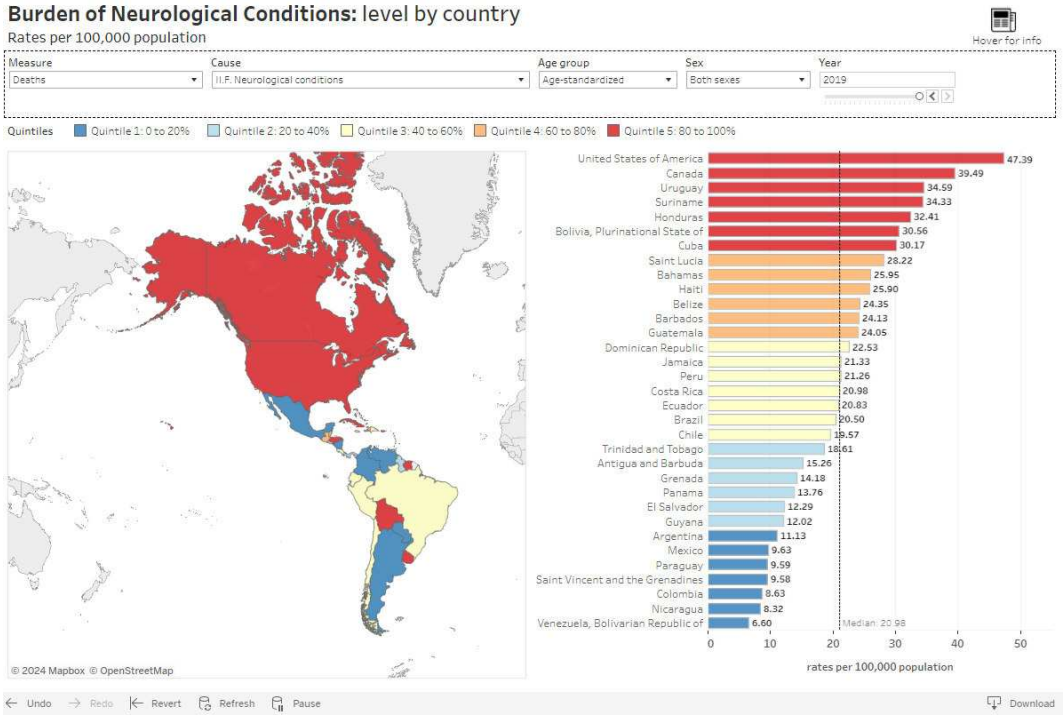
como YLLs) mostra que, de aproximadamente 7.5 milhões *YLLs*, 3.9 milhões foram de mulheres. Ainda, considerando grupos de 100 mil habitantes, com a padronização ponderada, para um total de 740 *YLLs*, 770 *YLLs* foram para as mulheres e 709 *YLLs* para os homens, e, com a padronização feita por idade, para um total de 552 *YLLs*, 604.4 *YLLs* foram para os homens e 499.1 *YLLs* foram para as mulheres.

Por fim, outro relevante parâmetro explorado foi o de anos vividos com a incapacidade, referido como *YLDs*, revelando que, de 8.2 milhões *YLDs*, 5.1 milhões *YLDs* foram de mulheres. Para grupos de 100 mil habitantes, com a padronização ponderada, para um total de 815.8 *YLDs*, 995.2 *YLDs* foram para as mulheres e 631.3 *YLDs* foram para os homens, e, com a padronização por idade, para um total de 737.3 *YLDs*, 880.3 *YLDs* foram para as mulheres e 589.4 *YLDs* foram para os homens, tendo o Brasil como o país com o maior nível.

Esses dados da OPAS são encontrados no estudo **The burden of Neurological conditions in the Region of the Americas** (PAHO, 2021) e, resumidamente, podem ser visualizados em www.paho.org/en/enlace/burden-neurological-conditions onde foram também disponibilizados:

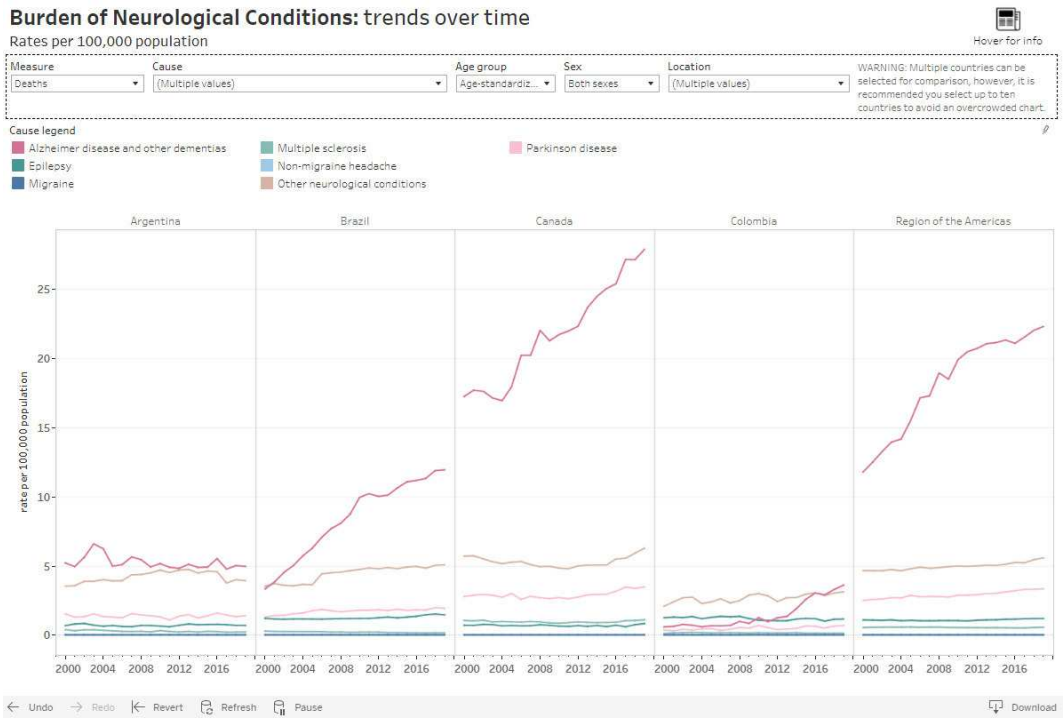
- um mapa com um painel de controle interativo (Figura 1), no qual se pode compreender dinamicamente o nível dos distúrbios nos países envolvidos;
- as tendências das doenças por país com o passar dos anos (Figura 2);
- as estatísticas da distribuição entre os países por sub-região (Figura 3);
- as estatísticas da distribuição por idade, sexo e causa (Figura 4).

Figura 1 – Mapeamento de doenças neurológicas nos países com painel interativo.



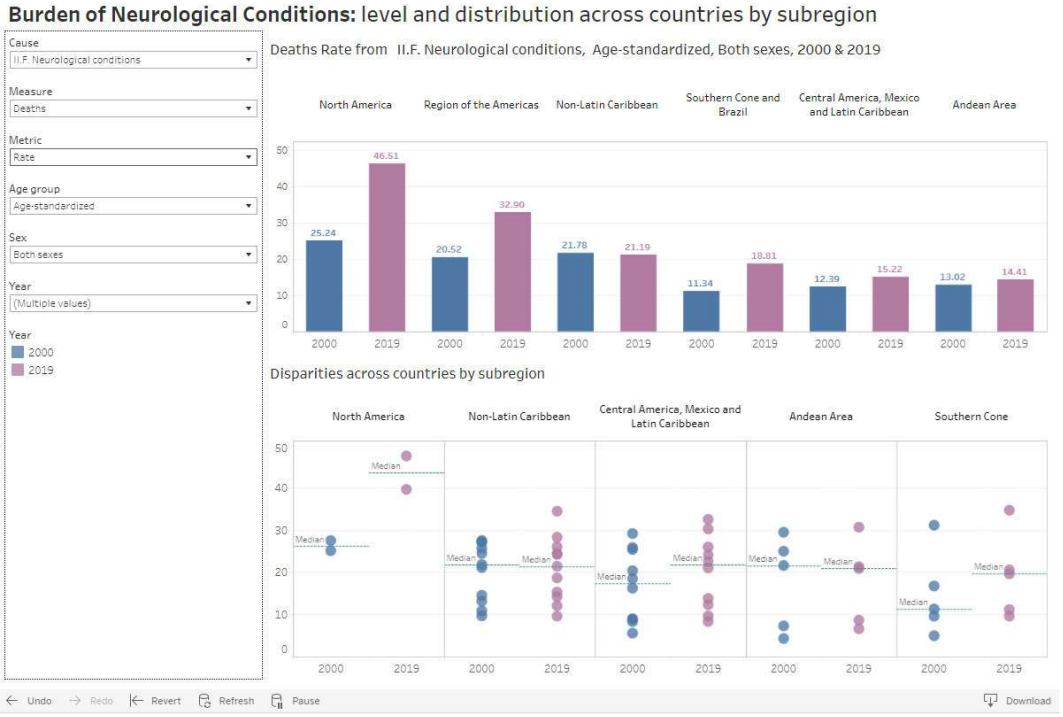
Fonte: https://www.paho.org/en/enlace/burden-neurological-conditions.

Figura 2 – Gráficos mostrando a tendência das doenças por país com o passar dos anos.



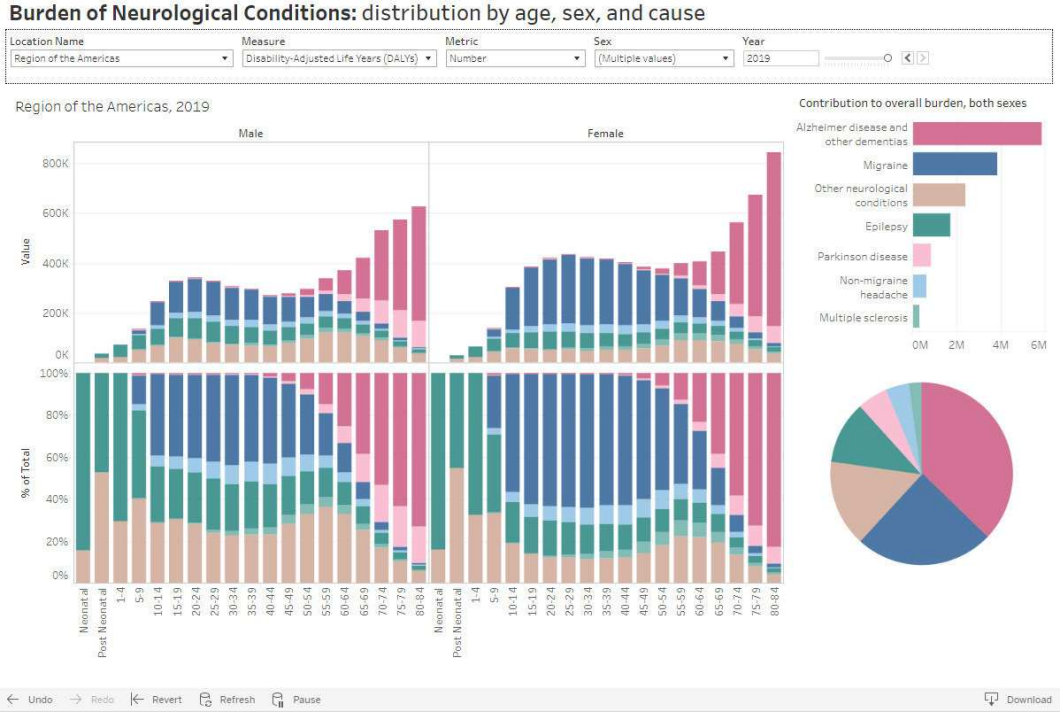
Fonte: https://www.paho.org/en/enlace/burden-neurological-conditions.

Figura 3 – Estatísticas de distribuição entre países por sub-região.



Fonte: https://www.paho.org/en/enlace/burden-neurological-conditions.

Figura 4 – Estatísticas de distribuição por idade sexo e causa.

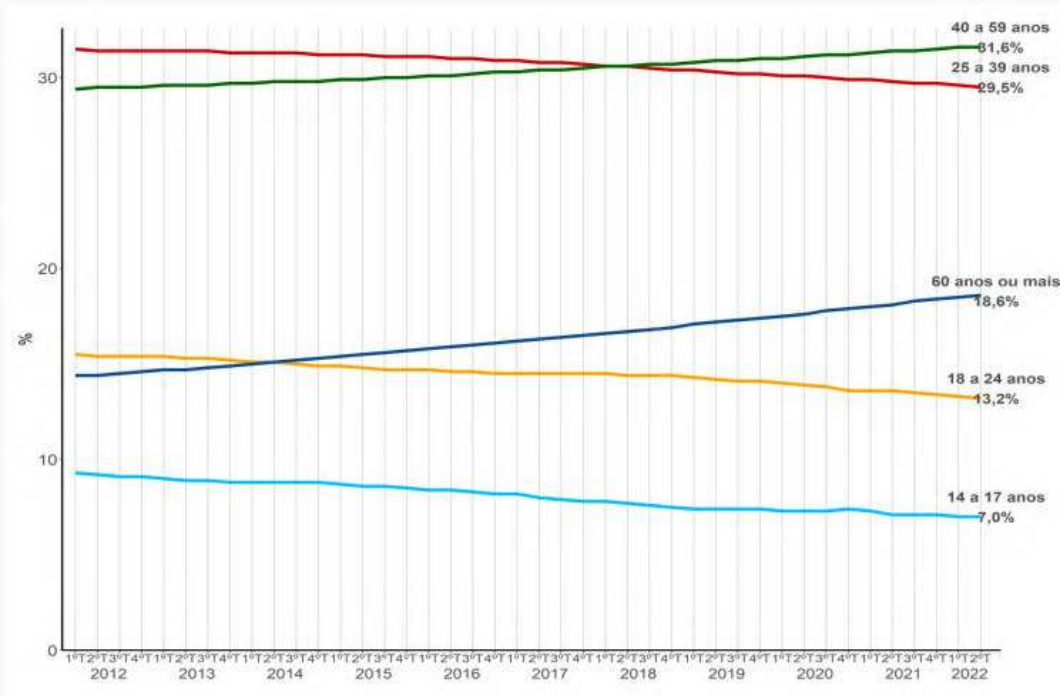


Fonte: https://www.paho.org/en/enlace/burden-neurological-conditions.

No cenário brasileiro, dado o aumento da expectativa de vida (IBGE, 2023), mostrado na Figura 5, a quantidade de idosos que se encontram na faixa etária mais afetada por condições

como *Alzheimer* e *Parkinson* está crescendo. Além disso, dietas inadequadas, sedentarismo e estresse urbano contribuem para a manifestação dessas doenças. (BVSMS, 2023a)

Figura 5 – Projeção da distribuição da população de 14 anos ou mais, por grupos de idade.



Fonte: https://agenciadenoticias.ibge.gov.br/arquivos/db973ee2b450d2303b0d3e622c67645b.pdf.

Outro relatório, produzido pela psiquiatra Cleusa Ferri, da Universidade Federal de São Paulo, destacou que pelo menos 1.76 milhão de pessoas com mais de 60 anos têm alguma forma de demência no Brasil. Além disso, o estudo alertou que cerca de 70% ou mais dos afetados não têm diagnóstico, e, portanto, não recebem tratamento adequado, prejudicando a qualidade de vida de modo geral. (BVSMS, 2023b)

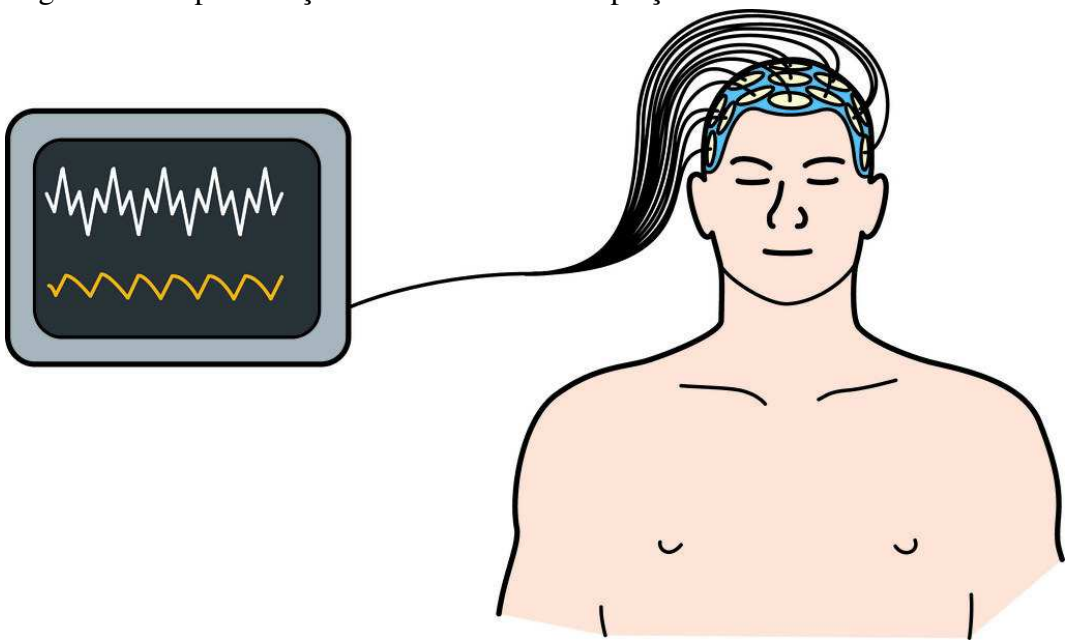
Como visto nesse contexto, as doenças neurológicas representam um desafio significativo para a saúde pública global, afetando milhões de pessoas e resultando em consideráveis impactos sociais e econômicos. Além das condições já mencionadas, quadros de epilepsia, esclerose múltipla, depressão, esquizofrenia e outras, não só comprometem a qualidade de vida dos pacientes, mas também, impõem uma carga pesada sobre família, amigos, quando estes não se afastam. Tais pacientes requerem uma atenção multiprofissional, incluindo cuidadores, quando há condições financeiras suficientes para contratar essa estrutura de apoio.

Uma das principais dificuldades no tratamento dessas e de outras doenças que envolvem o cérebro reside na complexidade de seus diagnósticos e na variabilidade dos sintomas apresentados por cada paciente. Diagnósticos precisos e precoces são cruciais para a imple-

mentação de tratamentos eficazes, mas, muitas vezes, os métodos tradicionais de diagnóstico clínico são insuficientes para captar a dinâmica completa da atividade cerebral subjacente a essas condições. Desse modo, a necessidade de técnicas avançadas e mais precisas torna-se evidente quando se considera a complexidade das origens, manifestações e tratamentos dessas doenças.

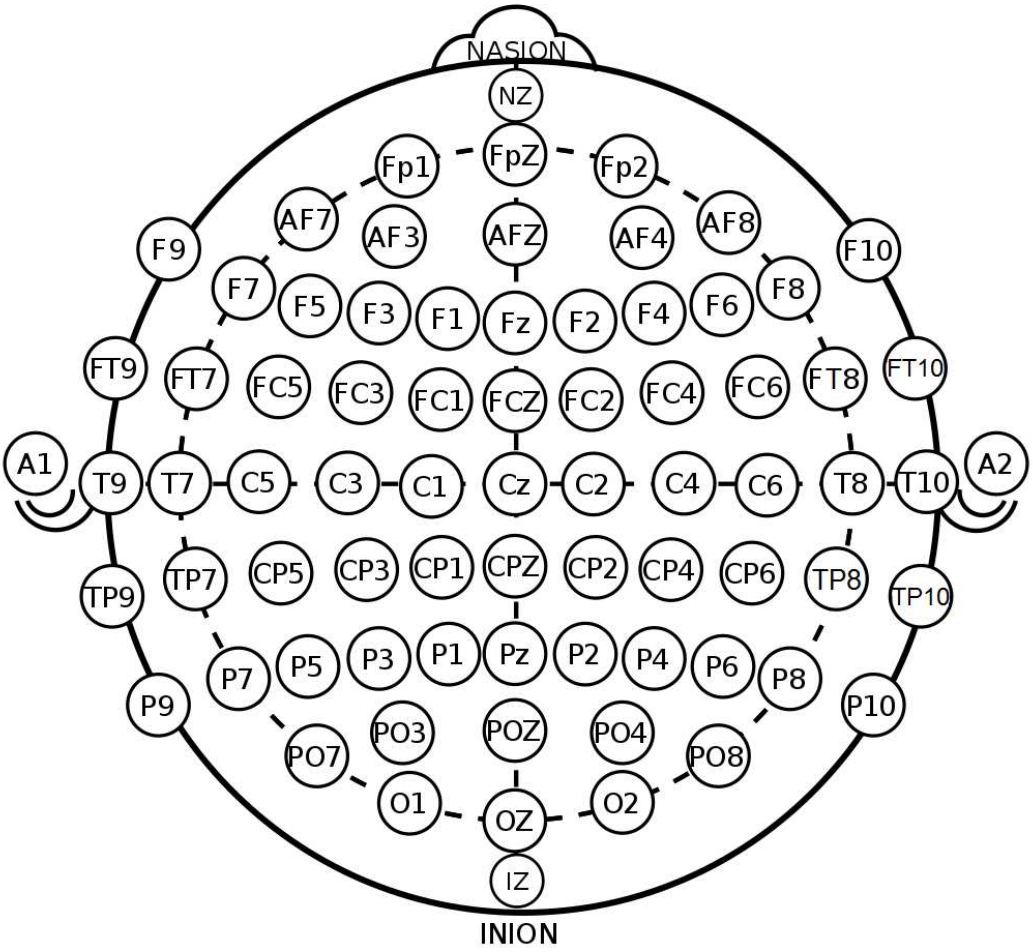
Nesse contexto, a Eletroencefalograma (EEG) se destaca como uma ferramenta essencial na avaliação da atividade elétrica do cérebro. O EEG é uma técnica não invasiva que registra as flutuações de tensão no couro cabeludo geradas pela atividade elétrica do cérebro usando uma touca, como representado na Figura 6, com eletrodos distribuídos de uma forma específica (Figura 7). Estas medições oferecem uma visão em tempo real da função cerebral, permitindo a detecção de anomalias que podem estar associadas a diversas doenças neurológicas. Assim, a capacidade de fornecer dados dinâmicos sobre o funcionamento cerebral torna o EEG indispensável na prática clínica e na pesquisa neurocientífica. (SUN; MOU, 2023)

Figura 6 – Representação do uso da touca e captação do sinal cerebral elétrico.



Fonte: Imagem para livre uso não comercial obtida em https://www.flickr.com/.

Figura 7 – Representação das posições dos eletrodos em uma touca.



Fonte: Imagem para livre uso não comercial obtida em https://commons.wikimedia.org/.

O EEG é útil para monitorar e analisar a atividade elétrica cerebral em diversas condições, como por exemplo, o Alzheimer, ajudando a detectar mudanças sutis na função cerebral antes mesmo do surgimento dos sintomas mais evidentes, a esquizofrenia e a depressão, auxiliando a identificar os padrões de atividade cerebral que podem ser correlacionados com estados emocionais e cognitivos alterados, a epilepsia, permitindo um diagnóstico preciso e a implementação de tratamentos adequados ao ajudar na localização de focos epilépticos. Além disso, o EEG tem aplicações importantes em outras doenças, como a esclerose múltipla, em que pode detectar anormalidades correspondentes a surtos da doença, ajudando a monitorar a progressão e a eficácia do tratamento. Em casos de lesões cerebrais traumáticas, ele é crucial para avaliar o grau de comprometimento e prever a recuperação. Ainda, em doenças neurodegenerativas raras, como a doença de *Huntington*, a técnica pode revelar padrões indicativos do estágio da doença, auxiliando no processo clínico investigativo. (FATIMA *et al.*, 2022)

Apesar de suas vantagens, a interpretação dos dados de EEG pode ser complexa

devido à enorme quantidade de informações geradas e à variabilidade individual dos sinais. Esta complexidade ocorre pela necessidade de distinguir entre atividades cerebrais normais e patológicas, o que requer um nível elevado de especialização e experiência. Além disso, os sinais de EEG estão frequentemente contaminados por ruídos, como movimentos oculares, musculares e artefatos provenientes de interferências externas. Esses ruídos podem dificultar a análise precisa dos dados e exigir técnicas avançadas de processamento para a sua remoção. A colocação dos eletrodos na touca de EEG também pode influenciar a qualidade do sinal, por isso, é necessária uma preparação cuidadosa e padronizada para garantir a consistência dos resultados. Em vista disso, a interpretação dos exames de EEG não só demanda conhecimento técnico detalhado sobre os padrões normais e anormais da atividade cerebral, mas também, habilidades na aplicação e interpretação de métodos de filtragem e análise de dados. Portanto, a pesquisa contínua e o desenvolvimento de novas técnicas de análise são essenciais para aprimorar a precisão diagnóstica e expandir as aplicações clínicas do EEG. (S, 2019)

Mais recentemente, os avanços nos algoritmos de aprendizagem de máquina têm mostrado grande potencial na análise de dados de EEG. As novas técnicas permitem processar grandes volumes de dados de maneira eficiente, identificando padrões sutis e complexos que podem passar despercebidos por métodos tradicionais e permitindo, com isso, uma análise detalhada, o que leva a diagnósticos mais precisos e personalizados.

Os algoritmos de classificação desempenham um papel crucial na categorização dos dados de EEG em diferentes classes, diferenciando doenças neurológicas de estados saudáveis. Dentre os principais algoritmos de classificação, destacam-se:

1. **Redes Neurais Artificiais (ANNs)**: Inspiradas na estrutura dos neurônios, são eficazes em reconhecer padrões não lineares nos dados de EEG. Estas redes também são amplamente utilizadas para classificar diferentes estados cerebrais, possibilitando uma análise detalhada da atividade elétrica do cérebro (MA *et al.*, 2021);
2. **Redes Neurais Convolucionais (CNNs)**: Capazes de identificar padrões multidimensionais devido ao uso de processamento matricial, que podem ser associados a diferentes condições neurológicas. Este método tem sido bastante útil na detecção de epilepsia, em que a precisão na identificação de focos epilépticos é crucial para o tratamento adequado (CHEN *et al.*, 2023);
3. **Máquinas de Vetores de Suporte (SVMs)**: As SVMs são eficazes em distinguir entre sinais normais e anormais, como na detecção de episódios epilépticos ou na identificação

de padrões associados a doenças psiquiátricas, como esquizofrenia e depressão. Esse algoritmo é especialmente útil em problemas de alta dimensionalidade, proporcionando diagnósticos precisos e confiáveis (ANTONY *et al.*, 2022);

4. **Florestas Aleatórias (*Random Forests*)**: Utilizam múltiplas árvores de decisão para melhorar a precisão da classificação, identificando características importantes nos sinais elétricos do cérebro associadas a diversas condições neurológicas, podendo, por exemplo, ser utilizadas para prever a progressão de doenças como Alzheimer. (YU *et al.*, 2024)

Há também os algoritmos de clusterização, que são fundamentais na análise de dados de EEG, pois ajudam a identificar padrões e agrupar dados similares sem a necessidade de rótulos pré-definidos:

1. **K-Médias (K-Means)**: Agrupa os dados em K *clusters*, onde cada ponto pertence ao *cluster* com o centroide mais próximo. No contexto de EEG, o *K-means* ajuda a identificar estados cerebrais distintos, detectando alterações abruptas na atividade cerebral (WEN; ARIS, 2022);
2. **Clusterização espacial baseada em densidade de dados com ruído (DBSCAN)**: Utilizado para identificar *clusters* de forma não linear com base na densidade dos pontos de dados. O DBSCAN é particularmente eficaz em lidar com dados de EEG ruidosos e pode detectar *clusters* de diferentes formas e tamanhos, além de identificar *outliers* que podem ser indicativos de eventos anômalos ou raros na atividade cerebral. Este algoritmo é útil para detectar surtos epilépticos e outras anomalias neurológicas (DU *et al.*, 2024);
3. **Floresta de Caminhos Ótimos (OPF)**: Utilizando princípios de aprendizado não supervisionado, busca encontrar agrupamentos naturais nos dados, representando eficientemente as relações e as distâncias entre os padrões identificados nos sinais elétricos cerebrais. Essa abordagem se mostra relevante na segmentação de diferentes estados ou eventos no EEG, oferecendo uma maneira robusta de diferenciar entre atividades cerebrais normais e anormais, contribuindo assim, para uma compreensão mais profunda das dinâmicas cerebrais e potencialmente para diagnósticos neurológicos mais precisos (NUNES *et al.*, 2014).

Assim, vê-se que a combinação do EEG, juntamente a técnicas avançadas de processamento de dados com a aplicação de algoritmos de aprendizagem de máquina, representa um grande avanço na forma como compreendemos e tratamos doenças neurológicas, abrindo novas fronteiras para a pesquisa e a prática clínica.

Esse trabalho compreende uma análise detalhada do uso de diversos algoritmos de clusterização em uma investigação do número ideal de *clusters* necessários para fazer a distinção adequada de indivíduos saudáveis daqueles que apresentam algum tipo de distúrbio neurológico. Seguindo o fluxo, a validação da conclusão majoritária é a submissão do conjunto de dados de EEG usado na etapa anterior (clusterização) para uma sequência de algoritmos de classificação. A partir disso, será feita a síntese dos resultados quantitativos e das percepções concluídas de forma a indicar as melhores escolhas envolvendo número de *clusters*, melhores algoritmos de clusterização e de classificação, assim como a configuração apropriada para cada um com base em diversas métricas de qualidade.

1.2 Objetivos

A intenção de desenvolver esse estudo parte da necessidade de investigar o número ótimo de grupos no processo de clusterização, especialmente para a obtenção dos microestados, termo que será explicado na fundamentação teórica, além de buscar entender como melhor utilizar alguns algoritmos de aprendizagem de máquina disponíveis na literatura para apoiar o diagnóstico clínico tradicional, fornecendo assim uma maior segurança aos profissionais da saúde e maior confiança dos pacientes no processo de investigação da doença.

A partir da pesquisa de fundamentação, os seguintes objetivos específicos foram pretendidos com a execução do trabalho:

- Verificar se o número de *clusters* usualmente empregado nos experimentos dos trabalhos acadêmicos é o mais adequado para se obter os melhores resultados;
- Testar se a maioria dos algoritmos de clusterização convergirá para o mesmo número de *clusters*;
- Examinar os valores das métricas de avaliação dos algoritmos de clusterização e de classificação frente ao possível número ótimo de *clusters*;
- Investigar a convergência do número ótimo de *clusters* necessários para fazer a separação clara entre os padrões elétricos associados a um cérebro saudável e a outro que apresenta uma doença;
- Disponibilizar um trabalho técnico validado para o uso dos algoritmos de clusterização K-Means, K-Medoids, Modified K-Means, AAHC e DBSCAN no contexto de EEG.

2 FUNDAMENTAÇÃO TEÓRICA

Nessa seção, serão expostas as informações necessárias para a compreensão de aspectos básicos relacionados ao contexto do trabalho, tais como a fisiologia e neurociência do cérebro, introdução ao EEG, microestados, doenças neurológicas, técnicas de análise e aplicações clínicas.

2.1 Fisiologia e Neurociência do cérebro

2.1.1 Geração de potenciais elétricos

A geração de potenciais elétricos no cérebro inicia no metabolismo energético das células, em especial, os neurônios. O metabolismo energético é o processo geral pelo qual as células obtém energia devido à quebra das ligações químicas das moléculas e, com reações adjacentes, realizam a oxidação dos nutrientes e a síntese de compostos de alta energia, especialmente a Adenosina Tri-Fosfato (ATP) que atua como o principal transportador de energia química em todas as células.

A partir dos mecanismos da síntese de ATP, os elétrons removidos pelo processo de oxidação das moléculas são transferidos para duas coenzimas transportadoras de elétrons, o NAD+ e o FAD, que são convertidas em suas formas reduzidas, NAD_H e FAD_{H_2}. Ainda, esse transporte de elétrons está associado ao bombeamento de prótons (H+) da matriz mitocondrial para o espaço entre as membranas celulares, gerando a força motriz dos prótons, uma combinação do gradiente de pH e do potencial elétrico na superfície da membrana. (CUENOUD *et al.*, 2020)

Diante disso, os potenciais elétricos do cérebro resultam da abertura e fechamento induzidos por despolarização das membranas dos canais iônicos, permitindo o fluxo passivo de íons Na+ e K+ para dentro e fora dos neurônios gerando um gradiente variável. Essa movimentação dos íons altera o potencial elétrico, gerando as correntes que se propagam para toda a estrutura do órgão em forma de ondas. (POIAN *et al.*, 2010)

2.1.2 Ondas cerebrais (Delta, Theta, Alpha, Beta)

A diferença entre os diversos tipos de ondas cerebrais existentes reside principalmente na faixa de frequência em que cada um está compreendido. Com o advento de técnicas de EEG mais avançadas, sobretudo por apresentar melhora na precisão da detecção do sinal,

foi possível desenvolver estudos os quais concluíram que cada tipo de onda está associado a diferentes estados da atividade mental, emocional e também física do indivíduo. (XAVIER *et al.*, 2020)

Os principais tipos de ondas cerebrais, assim como suas respectivas faixas de frequência e características, foram resumidos na Tabela 1.

Tabela 1 – Tipos de Ondas Cerebrais e Suas Características.

Tipo de Onda	**Frequência (Hz)**	**Características**
Ondas *Delta*	< 4	Ondas de baixa frequência observadas durante o sono profundo e restaurador, caracterizado por uma atividade cerebral lenta e sincronizada. São essenciais para a regeneração física e mental.
Ondas *Theta*	4 - 8	Ondas que surgem durante estados de relaxamento profundo, meditação leve, ou períodos de transição entre vigília e sono. Também podem estar presentes durante o sono REM.
Ondas *Alpha*	8 - 13	Ondas de atividade cerebral predominantes quando uma pessoa está relaxada e mentalmente calma, com os olhos fechados. Podem ser observadas durante a meditação ou estados de relaxamento profundo.
Ondas *Beta*	14 - 30	Ondas associadas ao estado de vigília e atividade mental concentrada. São mais proeminentes quando a pessoa está alerta, pensando ativamente ou envolvida em atividades que requerem foco e atenção.

Fonte: Criada pelo autor, baseando-se nas informações de (IDRIS *et al.*, 2024).

2.1.3 Relação entre atividade elétrica e funções cerebrais

Com o desenvolvimento do exame eletroencefalografia, cuja retrospectiva histórica será abordada na seção História e desenvolvimento do EEG, foi possível identificar que determinadas atividades humanas estão associadas a padrões elétricos específicos observados ao longo do tempo em diversos experimentos. Por exemplo, a análise do estado de repouso revela características distintas nos sinais elétricos do cérebro, sendo útil também para comparar um corpo descansando de um realizando alguma ação, (LIU *et al.*, 2024). Com isso, a execução de diversas tarefas pode ser monitorada e reconhecida através dos padrões elétricos correspondentes (HUSSAIN *et al.*, 2023). Outro campo de estudo que se beneficia do EEG é o reconhecimento de emoções onde mudanças nos sinais elétricos do cérebro refletem diferentes estados emocionais. (LIU *et al.*, 2021)

Além disso, mais alinhado ao contexto desse trabalho, destaca-se o uso do referido exame para a detecção e tratamento de doenças neurológicas. Dentre os trabalhos feitos,

podem ser citados artigos que relatam a identificação de Parkinson, (OBAYYA *et al.*, 2023), monitoramento de pacientes após sofrerem AVC, (SETIAWAN *et al.*, 2020), antecipação de ataques epilépticos, (QUYEN *et al.*, 2001), e outros para demais doenças.

2.2 Introdução ao EEG

2.2.1 História e desenvolvimento do EEG

A descoberta da eletroencefalografia (EEG), em 1929, pelo psiquiatra alemão Hans Berger foi um marco histórico, fornecendo uma nova ferramenta de diagnóstico neurológico e psiquiátrico na época. Antes disso, muitos outros cientistas interagiram com o aspecto elétrico do cérebro, a citar, Rolando, que capturou estímulos elétricos no topo da cabeça e, com os resultados disso, Fritsch e Hitzig, desenvolvendo posteriormente a ideia de localização cerebral, ou seja, partes do cérebro estão relacionadas a partes do corpo, uma vez que o estímulo elétrico em diferentes locais causava movimentações em diferentes membros. Porém, com os estudos desenvolvidos, não se podia concluir que havia uma atividade elétrica natural que pudesse ser detectada.

Caton, usando um cérebro aberto extraído de um corpo, foi o primeiro a relatar corrente elétrica na substância cinzenta. Inspirado com a descoberta, Berger, empregando um amplificador comum de rádio e um papel, fez a gravação de um eletroencefalograma ao participar de uma operação neurocirúrgica, considerando as correntes como artefatos, mas nomeando-as com os termos ondas alfa e ondas beta. (TUDOR *et al.*, 2005)

Offner desenvolveu um equipamento para EEG que usava eletrodos de agulha, portanto, invasivo ao corpo, mas que possibilitou Gibbs, em 1935, a descrever as características das ondas cerebrais e, após isso, junto a Jasper, mostraram os inerentes picos elétricos que ocorrem entre as crises de uma pessoa com epilepsia. Após a Segunda Guerra Mundial, os pesquisadores se empenharam em construir melhores dispositivos eletrônicos, especialmente os amplificadores, e também desenvolver diferentes métodos de detecção, limpeza e classificação dos sinais cerebrais de forma a diferir sinais anormais dos sinais de um cérebro considerado saudável. Por fim, Walter criou a técnica EEG topográfico, o que permitiu um mapeamento geral da atividade elétrica de toda a superfície do cérebro. (AL-KADI *et al.*, 2013)

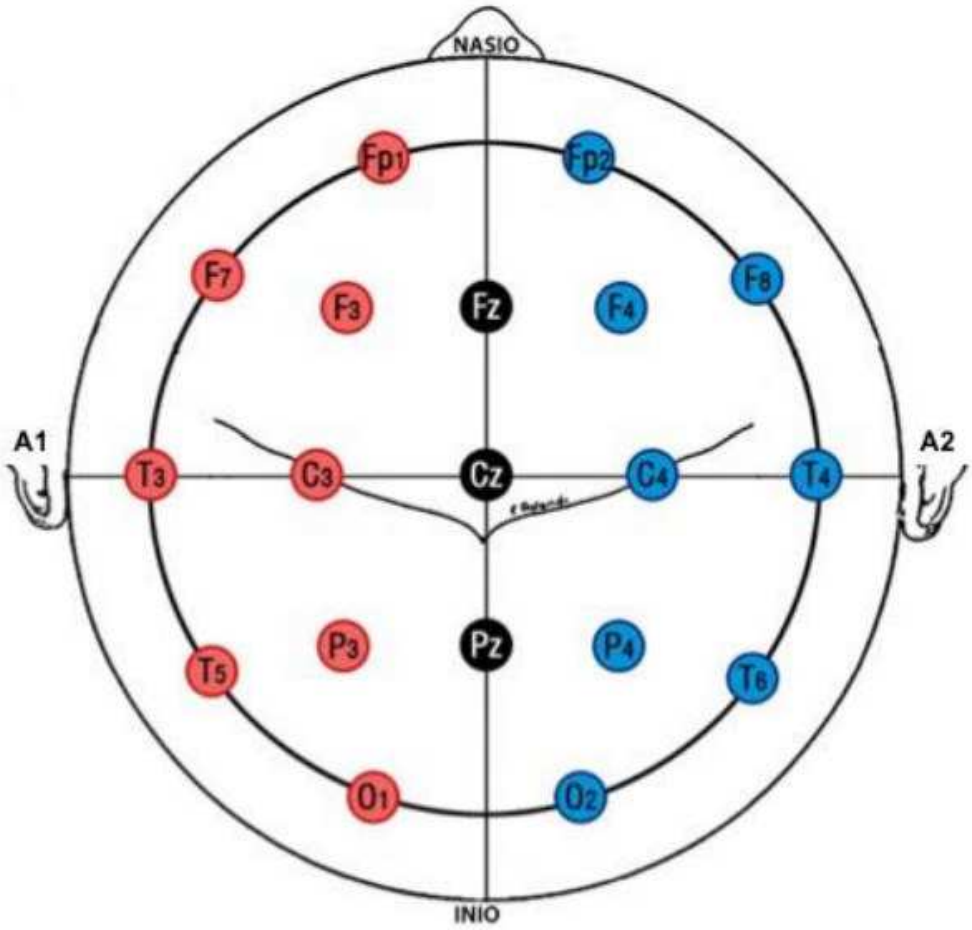
2.2.2 Princípios básicos de funcionamento do EEG

O eletroencefalograma funciona por meio da colocação de pequenos eletrodos junto a um gel condutivo no couro cabeludo que captam as variações elétricas geradas pela atividade neuronal. Cada eletrodo detecta sinais elétricos provenientes de uma região da cabeça, permitindo a medição das diferenças do potencial elétrico entre diferentes pontos do cérebro. Essas diferenças, por serem de baixa magnitude e ocorrerem múltiplas vezes em menos de um segundo, precisam ser amplificadas, filtradas devido à existência de ruídos e registradas como ondas em um gráfico para que seja realizada a análise clínica por um profissional da saúde adequado. (LIGHT *et al.*, 2010)

Mais detalhadamente, o processo para realização do exame inicia com a tranquilização do paciente e a preparação do topo da cabeça para o posicionamento adequado dos eletrodos, geralmente com o suporte de uma touca apropriada e aplicação de gel que facilite a captação elétrica. Ainda sobre o sistema de aquisição, o sinal capturado é amplificado por meio de um amplificador eletrônico, cujo sinal resultante sofrerá filtragens para eliminar possíveis ruídos inerentes e chegará ao dispositivo principal, o qual, dependendo da programação, irá aplicar mais amplificações ou filtragens a nível de *software*, e mostrará o resultado em uma tela ou fará a impressão do registro gráfico ao longo do tempo de captação. (PINEGGER *et al.*, 2016)

Outro ponto a se destacar é a quantidade e os locais onde os eletrodos são posicionados, os quais influenciam na obtenção da atividade elétrica durante o exame. O padrão mais encontrado na literatura e empregado em diferentes países é o sistema 10-20 (mostrado na Figura 8) apresentado por H. H. Jasper em 1957, onde, dos 21 eletrodos, 19 são dispostos no topo do crânio e os outros dois são colocados nos lóbulos das orelhas (marcações A1 e A2 na figura), ambos servindo como pontos de referência para as demais captações elétricas. Ainda, na Figura 8, estão destacados os pontos craniométricos NASIO, referenciando o nariz, e INIO, referenciando a região posterior da cabeça). (KLEM *et al.*, 1999)

Figura 8 – Posicionamento dos eletrodos no crânio (Sistema 10-20).



Fonte: Sociedade Brasileira de Neurofisiologia Clínica. Recomendação da SBNC para Localização de Eletrodos e Montagens de EEG. São Paulo: SBNC, 27 de novembro de 2017.

Ainda, os ruídos, como são chamados os sinais que não fazem parte do domínio de interesse, que são inerentes ao exame, podem ser classificados como artefatos intrísecos e extrínsecos ao corpo, os primeiros sendo de origem fisiológica (atividade muscular, batimento cardíaco, piscar dos olhos, entre outros), e os demais de origem externa (movimentação dos eletrodos ou do fio de condução até o dispositivo principal, defeito na circuitaria envolvida, entre outros). Tal presença gera a necessidade de uma etapa adequada na filtragem para que o sinal advindo da atividade neuronal possa ser avaliado corretamente. (JIANG *et al.*, 2019)

2.3 Microestados

Esta seção apresentará o conceito de microestados em um eletroencefalograma, que são fundamentais para a compreensão da motivação dos experimentos e importantes como biomarcadores na detecção de múltiplas doenças neurológicas. Além disso, serão abordadas os

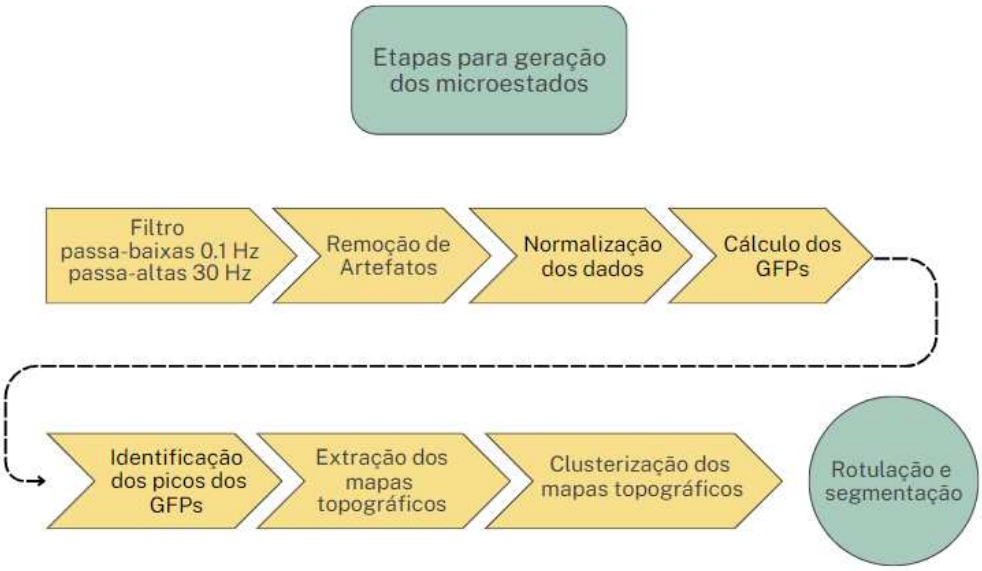
diferentes algoritmos de aprendizagem de máquina.

2.3.1 Definição de microestados

O artigo de origem do conceito de microestados foi proposto por Lehmann e outros colaboradores que analisaram uma frequência beta de um sinal EEG e conseguiram decompô-la em um número limitado de estados quase estáveis (LEHMANN *et al.*, 1987). Formalmente, microestados são registros topográficos dos potenciais elétricos que ficam estáveis por 80 a 120 milissegundos antes de ocorrer a transição para outros microestados. (KHANNA *et al.*, 2015)

A Figura 9 apresenta o fluxograma das etapas envolvidas na geração dos microestados a partir de um exame EEG.

Figura 9 – Etapas para geração dos microestados.



Fonte: Elaborada pelo autor.

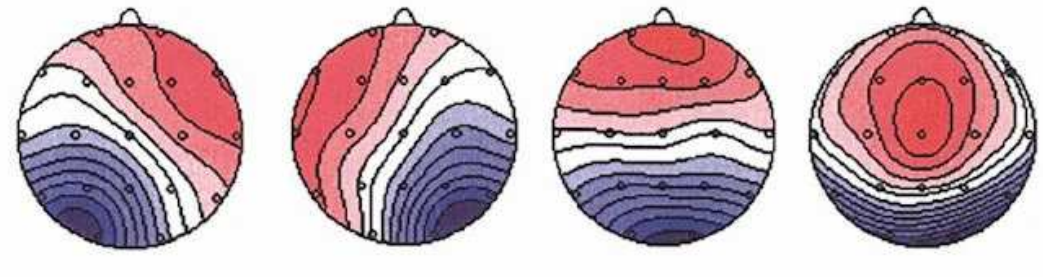
Com o emprego de bases de dados variadas para realizar comparações entre indivíduos de variados grupos, diversos estudos mostram a distinção entre as características dos microestados para comportamentos divergentes (STEVENS; KIRCHER, 1998), personalidades (SCHLEGEL *et al.*, 2012) e doenças neurológicas (ZOUBI *et al.*, 2019; IFTIMOVICI *et al.*, 2023). Partindo desses e de outros estudos, pode-se inferir que diferentes tipos de microestados estão associados a diferentes estados cognitivos no contexto das doenças cerebrais, já que, por premissa, pessoas doentes correspondem a diferentes registros de microestados em comparação a pessoas saudáveis.

Os quatro microestados mais comuns, conforme descrito em (TARAILIS *et al.*,

2024), estão representados na Figura 10. Esta figura apresenta visões superiores do crânio, com o nariz do indivíduo fora dos círculos principais. As linhas indicam níveis de mesmo potencial elétrico, enquanto as cores mostram a atividade elétrica: vermelho para maior atividade positiva e azul para maior atividade negativa, considerando uma normalização dos sinais no intervalo [-1, 1]. Na literatura, os microestados são frequentemente associados a padrões específicos de estado mental e funções cognitivas, caracterizando-se da seguinte forma:

- **Microestado A**: apresenta-se em uma configuração frontal direita para posterior esquerda, ocorrendo em maior quantidade nos experimentos que envolvem ativação no córtex temporal e ativação da rede auditiva, embora existam trabalhos que também encontraram correlação desse microestado com a rede visual (ANTONOVA *et al.*, 2022);
- **Microestado B**: apresenta-se em uma configuração frontal esquerda para posterior direita, ocorrendo significativamente em atividades de processamento visual, principalmente em ações que possuem exposição a ambientes visuais com muitos elementos, a citar jogos de videogame. Ainda, é também encontrado em atividades que envolvam autovisualização, memórias autobiográficas (CUI *et al.*, 2021);
- **Microestado C**: apresenta-se em uma configuração simétrica anterior para posterior, ocorrendo em atividades que envolvem relaxamento e autorreflexão (TOMESCU *et al.*, 2022);
- **Microestado D**: apresenta-se em uma configuração fronto-central e está mais presente em atividades que exigem controle cognitivo e atenção plena (MURPHY *et al.*, 2018).

Figura 10 – As quastro topografias dos microestados mais comuns na literatura.



Fonte: Disponível em https://brainlatam.com.

2.4 Algoritmos de aprendizagem de máquina

Esta seção apresentará em detalhes os algoritmos usados tanto na etapa de clusterização quanto na de classificação, explicitando também as melhores situações para usar cada um

deles.

2.4.1 Algoritmos de clusterização

Algoritmos de clusterização são técnicas de aprendizado de máquina não supervisionado que agrupam dados em subconjuntos chamados *clusters*, sem a necessidade de rótulos ou classes previamente definidos. Eles são projetados para identificar padrões e estruturas nos dados, maximizando a semelhança dentro dos *clusters* e a diferença entre eles.

Uma abordagem comum na clusterização é agrupar dados com base na proximidade espacial. Isso significa que os dados são organizados de acordo com a distância entre eles em um espaço multidimensional, formando *clusters* de itens que estão mais próximos uns dos outros.

Outra técnica frequentemente usada envolve a densidade dos pontos de dados. Aqui, *clusters* são identificados em regiões onde os pontos de dados são densamente agrupados, diferenciando essas áreas de regiões com baixa densidade. Isso permite a identificação de grupos que podem variar em forma e tamanho, e que podem incluir áreas de ruído ou *outliers*, que são dados que se desviam significativamente da maioria dos pontos em um conjunto, indicando valores extremos ou anômalos que podem resultar de variabilidade natural, erros de medição ou condições incomuns.

A clusterização hierárquica é uma técnica que forma uma árvore de *clusters*, permitindo visualizar os dados em diferentes níveis de detalhe. Este método começa agrupando os pontos de dados mais próximos e, gradualmente, combina *clusters* menores em *clusters* maiores, oferecendo uma visão detalhada das relações entre os dados em vários níveis.

Outra abordagem utiliza representações em grafos dos dados, onde os pontos são representados como nós e as conexões entre eles são baseadas em similaridade ou proximidade. *Clusters* são formados agrupando nós que estão fortemente conectados, revelando comunidades ou grupos de dados que possuem uma forte interconexão.

Essas técnicas permitem a extração de informações valiosas e a identificação de padrões ocultos nos dados. Isso as torna ferramentas poderosas para a análise exploratória de dados em diversas áreas, como saúde e economia, revelando *insights* que poderiam passar despercebidos de outra forma. Essas ferramentas são fundamentais em contextos acadêmicos e profissionais, fornecendo uma base sólida para tomadas de decisão e avanços na pesquisa.

No contexto do trabalho, os seguintes algoritmos foram escolhidos para realizar a etapa de clusterização e os respectivos resultados usados para o cálculo das métricas relacionadas:

- **K-médias (em inglês, *K-Means*)**: segundo (IKOTUN *et al.*, 2023), o K-Médias é um algoritmo que divide um conjunto de dados em k grupos, ou *clusters*, de modo a minimizar a variabilidade interna de cada um. O processo inicia com a seleção aleatória de k centroides. Em seguida, o algoritmo atribui cada ponto de dado ao *cluster* cujo centróide está mais próximo, calculando a distância euclidiana:

$$\|x_j - \mu_i\| = \sqrt{\sum_{d=1}^{D} (x_{jd} - \mu_{id})^2} \tag{2.1}$$

onde x_j é o ponto de dado, μ_i é o centróide do *cluster* i, e D é o número de dimensões. Após a atribuição dos pontos, os centroides de cada *cluster* são recalculados pela média das posições dos pontos pertencentes ao *cluster*:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \tag{2.2}$$

Este processo é repetido até que os centroides não mudem significativamente. O objetivo é minimizar a função de custo, que representa a soma das distâncias quadráticas dos pontos ao centróide do seu *cluster*:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \tag{2.3}$$

- ***K-Medoids***: segundo (ARORA *et al.*, 2016), *K-Medoids* é um algoritmo semelhante ao *K-Means*, mas, em vez de usar a média dos pontos do *cluster* como centróide, ele usa um ponto intermediário real como *medoid*, considerado o representante do *cluster*. Este ponto é aquele cuja soma das distâncias para todos os outros pontos do *cluster* é mínima. O algoritmo é mais tolerante a *outliers* e ruídos do que o *K-Means*.

  No K-Medoids, o processo de clusterização começa com a seleção inicial de k medoids. Em seguida, cada ponto de dado é atribuído ao *cluster* cujo medoid está mais próximo, de acordo com a distância:

$$\|x_j - m_i\| = \min_{x_j \in C_i} \sum_{x_k \in C_i} \|x_j - x_k\| \tag{2.4}$$

onde x_j é o ponto de dado, m_i é o medoid do *cluster i*, e C_i é o conjunto de pontos pertencentes ao *cluster i*. O medoid m_i é o ponto que minimiza a soma das distâncias para todos os outros pontos do *cluster*.

Após a atribuição dos pontos, os medoids são recalculados. Para isso, o algoritmo verifica cada ponto do *cluster* para encontrar aquele cuja soma das distâncias aos demais pontos seja a menor possível, tornando-o o novo medoid:

$$m_i = \arg\min_{x_j \in C_i} \sum_{x_k \in C_i} \|x_j - x_k\| \tag{2.5}$$

Este processo é repetido até que os medoids não mudem mais significativamente ou até que a mudança na função de custo atinja um valor abaixo de um determinado limiar. A função de custo do K-Medoids é dada por:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - m_i\| \tag{2.6}$$

- **K-médias modificado (em inglês, *Modified K-Means*)**: segundo (PATEL; MEHTA, 2011), o K-médias modificado é uma variação do *K-Means* que inclui modificações para melhorar a performance ou a acurácia do algoritmo original. Essas modificações podem incluir inicializações aprimoradas de centroides, métodos diferentes para atualizar os centroides, ou a inclusão de pesos nos pontos de dados para influenciar a formação dos *clusters*.

  Uma dessas modificações comuns é a inicialização aprimorada dos centroides, como no método *K-Means++*, que seleciona centroides iniciais distantes entre si para evitar problemas de convergência para mínimos locais. Outra variação é a atualização ponderada dos centroides, onde a nova posição de um centróide é calculada levando em conta pesos associados a cada ponto de dado:

$$\mu_i = \frac{\sum_{x_j \in C_i} w_j x_j}{\sum_{x_j \in C_i} w_j} \tag{2.7}$$

  onde w_j é o peso atribuído ao ponto de dado x_j. A inclusão de pesos pode ajudar a ajustar o algoritmo para refletir a importância de certos pontos de dados, aumentando a adaptabilidade a diferentes tipos de dados.

  Além disso, algumas versões do *Modified K-Means* incluem estratégias para atualizar os centroides, como excluir outliers ou utilizar medidas de distância diferentes da euclidiana.

Tais modificações permitem que o algoritmo lide melhor com conjuntos de dados que apresentam variabilidade significativa ou formas de *clusters* não esféricos.

- **Aglomeração Hierárquica (em inglês, *Atomize and Agglomerative Hierarchical Clustering*, AAHC)**: é um algoritmo de clusterização hierárquica que começa tratando cada ponto de dado como um *cluster* individual. Em seguida, os *clusters* são combinados sucessivamente com os *clusters* mais próximos, até que todos os pontos pertençam a um único *cluster* ou que um número pré-determinado de *clusters* seja alcançado.

  A similaridade ou distância entre os *clusters* é geralmente medida pela distância euclidiana, mas também podem ser usadas outras medidas. O algoritmo pode ser formalizado da seguinte maneira:

  1. **Inicialização**: Cada ponto de dado é tratado como um *cluster* individual;
  2. **Combinação de *clusters***: Em cada iteração, os dois *clusters* mais próximos são combinados em um novo *cluster*. A distância entre dois *clusters* C_i e C_j é calculada, por exemplo, usando a ligação média:

     $$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x_p \in C_i} \sum_{x_q \in C_j} \|x_p - x_q\| \tag{2.8}$$

     onde $|C_i|$ e $|C_j|$ são os números de pontos nos *clusters* C_i e C_j, respectivamente, e $\|x_p - x_q\|$ é a distância entre os pontos x_p e x_q;
  3. **Repetição**: O processo de combinação continua até que um único *cluster* seja formado ou até que o número desejado de *clusters* seja alcançado.

  Este método hierárquico é particularmente útil para revelar a estrutura dos dados, como a formação de *clusters* em diferentes níveis de granularidade. A vantagem do AAHC é que ele não requer a definição prévia do número de *clusters*, ao contrário do K-Means, permitindo uma análise mais flexível e detalhada da distribuição dos dados. (SCHILLING; COGGINS, 2007);

- **Agrupamento baseado em densidade de dados com ruído (em inglês, *Density-Based Spatial Clustering of Applications with Noise*, DBSCAN)**: segundo (KUI *et al.*, 2023), o DBSCAN é um algoritmo de clusterização que busca identificar regiões de alta densidade de pontos de dados separadas por regiões de baixa densidade. Ele é eficaz na identificação de *clusters* com formas arbitrárias e na detecção de *outliers* como pontos que não pertencem a nenhum *cluster*. O DBSCAN exige dois parâmetros principais: a distância máxima entre pontos para que sejam considerados vizinhos (ε) e o número mínimo de pontos (*minPts*)

para formar um *cluster* denso. O processo de clusterização no DBSCAN é baseado nos seguintes conceitos:

1. **Ponto Central**: Um ponto p é considerado um ponto central se pelo menos existe *minPts* pontos dentro de sua vizinhança definida pelo raio ε;
2. **Ponto de Borda**: Um ponto que não é um ponto central, mas está na vizinhança (ε) de um ponto central;
3. ***Outlier* (ponto de ruído)**: Um ponto que não é um ponto central nem está na vizinhança de nenhum ponto central.

O algoritmo DBSCAN inicia percorrendo todos os pontos do conjunto de dados. Para cada ponto p, ele verifica se existem pelo menos *minPts* pontos dentro de uma distância ε. A vizinhança $N_\varepsilon(p)$ de um ponto p é definida por:

$$N_\varepsilon(p) = \{q \in D \mid \|p - q\| \leq \varepsilon\} \tag{2.9}$$

onde D é o conjunto de dados, e $\|p - q\|$ é a distância entre os pontos p e q.

Se o ponto p satisfaz as condições de ser um ponto central, um novo *cluster* é formado, e todos os pontos conectados a p dentro de ε são adicionados ao *cluster*. O processo é repetido para todos os pontos até que todos estejam atribuídos a um *cluster* ou marcados como *outliers*.

O DBSCAN é vantajoso por não exigir a especificação do número de *clusters* previamente e por sua habilidade de identificar *clusters* com formas variadas, além de detectar *outliers*. Contudo, sua eficácia depende fortemente da escolha adequada dos parâmetros ε e *minPts*.

2.4.2 Algoritmos de classificação

Algoritmos de classificação são técnicas de aprendizagem de máquina supervisionada, utilizadas para prever rótulos ou classes de novos dados baseando-se em exemplos previamente rotulados. Diferentemente dos algoritmos de clusterização, que agrupam dados sem rótulos, os algoritmos de classificação aprendem com dados rotulados para realizar previsões precisas.

Uma abordagem comum envolve métodos baseados em distância que classificam novos dados com base na similaridade a exemplos conhecidos, atribuindo o rótulo mais frequente entre os mais próximos.

Técnicas probabilísticas também são amplamente utilizadas, calculando a probabilidade de um dado pertencer a uma determinada classe e escolhendo a classe com a maior probabilidade, sendo úteis em situações com incerteza ou variabilidade nos dados.

Essas técnicas, essenciais para automatizar processos decisórios e análise de dados, possibilitam uma classificação eficiente e precisa, fornecendo uma base robusta para aplicações em diversas áreas, como diagnóstico médico, detecção de fraudes e reconhecimento de padrões.

- **Floresta Aleatória (em inglês, *Random Forest* (RF))**: segundo (ZIEGLER; KöNIG, 2014), Floresta Aleatória é um algoritmo que envolve a criação de múltiplas árvores de decisão, cada uma treinada com um subconjunto aleatório dos dados. Cada árvore emite uma "votação"para a classe do dado de entrada, e a classe final é determinada pela maioria dos votos, reduzindo o risco de *overfitting* e melhorando a precisão.

  Para problemas de classificação, a classe final $\hat{y}$ é dada pelo voto majoritário das árvores:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\} \tag{2.10}$$

  Para regressão, a previsão é a média das previsões das árvores:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} T_i(x) \tag{2.11}$$

  Durante o treinamento, a seleção de subconjuntos de dados é feita por *bootstrap sampling*, e os nós das árvores são divididos com base em critérios como o erro quadrático médio (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.12}$$

  onde N é o número de exemplos no conjunto de dados, y_i é o valor real do i-ésimo exemplo, e $\hat{y}_i$ é o valor previsto pela árvore para esse exemplo.

- **Máquinas de Vetores de Suporte (em inglês, *Support Vector Machine* (SVM))**: funciona encontrando o hiperplano que melhor separa as classes de dados no espaço de características. O objetivo é maximizar a margem entre as classes, melhorando a capacidade do modelo de generalizar para novos dados.

  A margem é definida como a distância entre o hiperplano separador e os vetores de suporte, que são os pontos de dados mais próximos do hiperplano. Matematicamente, o hiperplano em um espaço n-dimensional pode ser representado como:

$$w \cdot x + b = 0 \tag{2.13}$$

onde w é o vetor de pesos, x é o vetor de características, e b é o termo de viés.

O objetivo do SVM é encontrar os valores de w e b que maximizem a margem. A margem é dada por $\frac{2}{\|w\|}$, e para maximizar essa margem, o SVM resolve o seguinte problema de otimização:

$$\min \frac{1}{2}\|w\|^2 \tag{2.14}$$

sujeito à restrição:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i \tag{2.15}$$

onde y_i é a classe do ponto x_i (assumindo valores $+1$ ou -1). Esta formulação garante que os pontos de dados estejam corretamente classificados e fora da margem.

Quando os dados não são linearmente separáveis, o SVM pode utilizar uma função de custo com termos de penalização para violações à margem, ou aplicar funções de kernel para projetar os dados em um espaço de maior dimensão, tornando possível a separação (HEARST *et al.*, 1998);

- **Regressão Logística (em inglês, *Logistic Regression* (LR))**: estima a probabilidade de uma amostra pertencer a uma classe específica usando uma função logística. O modelo ajusta os coeficientes das características para minimizar a diferença entre as previsões e os valores reais, gerando uma fronteira de decisão linear.

  A função logística, ou *sigmoide*, é utilizada para mapear as previsões para um intervalo de 0 a 1, representando a probabilidade da amostra pertencer à classe positiva. A função é definida como:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.16}$$

onde $z = w \cdot x + b$, com w sendo o vetor de pesos, x o vetor de características, e b o termo de viés.

A probabilidade prevista de uma amostra x_i pertencer à classe positiva é dada por:

$$P(y_i = 1|x_i) = \sigma(w \cdot x_i + b) \tag{2.17}$$

Para ajustar os coeficientes w e b, a Regressão Logística utiliza a função de custo conhecida como *log-loss* ou entropia cruzada, que mede a diferença entre as previsões e os valores reais:

$$J(w,b) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\sigma(w \cdot x_i + b)) + (1 - y_i)\log(1 - \sigma(w \cdot x_i + b))] \tag{2.18}$$

onde N é o número de amostras, y_i é o valor real da classe da i-ésima amostra (0 ou 1), e $\sigma(w \cdot x_i + b)$ é a probabilidade prevista.

O objetivo é minimizar $J(w,b)$ para encontrar os melhores coeficientes que separam as classes, gerando uma fronteira de decisão linear no espaço das características (TOIVONEN *et al.*, 2019);

- **K-Vizinhos Próximos (em inglês, *K-Nearest Neighbors* (KNN))**: classifica um novo dado com base na classe mais comum entre seus k vizinhos mais próximos no espaço de características. É um método simples e eficaz que funciona bem com distribuições de dados não lineares.

  Para classificar um novo ponto x, o KNN calcula a distância entre x e todos os pontos do conjunto de treinamento. A distância mais comum é a distância Euclidiana, dada por:

$$d(x,x_i) = \sqrt{\sum_{j=1}^{n}(x_j - x_{ij})^2} \tag{2.19}$$

  onde x_j é a coordenada do novo ponto nas n dimensões, e x_{ij} é a coordenada do i-ésimo ponto de treinamento.

  Após calcular as distâncias, o algoritmo seleciona os k pontos mais próximos e determina a classe mais frequente entre eles. A classe do novo ponto x é então atribuída com base na votação majoritária dos k vizinhos:

$$\hat{y} = \text{mode}\{y_1, y_2, \ldots, y_k\} \tag{2.20}$$

  onde y_i representa a classe do i-ésimo vizinho mais próximo.

A escolha do parâmetro k é crucial para o desempenho do algoritmo: valores pequenos de k podem torná-lo sensível a ruídos nos dados, enquanto valores grandes podem suavizar a fronteira de decisão (UDDIN *et al.*, 2022);

- **Árvore de Decisão (em inglês, *Decision Tree* (DT))**: cria uma árvore onde cada nó representa uma decisão baseada em um atributo específico dos dados. A árvore é construída dividindo repetidamente o conjunto de dados em subgrupos homogêneos até que os nós terminais contenham instâncias predominantemente de uma única classe.

  Durante o processo de construção da árvore, são utilizadas métricas para selecionar o melhor atributo para dividir os dados em cada nó. Uma das métricas mais comuns é a entropia, que mede o grau de desordem ou impureza em um conjunto de dados:

$$H(S) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{2.21}$$

  onde S é o conjunto de dados, C é o número de classes, e p_i é a proporção de elementos da classe i no conjunto S.

  A partir da entropia, calcula-se o ganho de informação para determinar o melhor atributo para dividir o nó. O ganho de informação é a redução na entropia após dividir os dados com base em um atributo A:

$$IG(S,A) = H(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} H(S_v) \tag{2.22}$$

  onde $\text{Valores}(A)$ são os possíveis valores do atributo A, S_v é o subconjunto de S para o qual o atributo A tem o valor v, e $|S_v|/|S|$ é a proporção do subconjunto em relação ao conjunto original.

  A árvore continua a se dividir até atingir uma condição de parada, como atingir uma profundidade máxima ou quando os nós se tornam puros (contêm instâncias de uma única classe) (SONG; LU, 2015);

- **Impulsão de Gradiente (em inglês, *Gradient Boosting* (GB))**: constrói um modelo forte combinando múltiplos modelos fracos (árvores de decisão), treinando cada nova árvore para corrigir os erros das anteriores. Esse método iterativo melhora continuamente a precisão da classificação.

  O processo começa com um modelo simples, como uma árvore de decisão rasa. A cada

iteração m, o objetivo é ajustar uma nova árvore aos resíduos $r_i^{(m)}$, que são as diferenças entre os valores reais y_i e as previsões anteriores $F_{m-1}(x_i)$:

$$r_i^{(m)} = y_i - F_{m-1}(x_i) \tag{2.23}$$

O modelo final $F_M(x)$ é a soma ponderada dos modelos ajustados em cada iteração:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x) \tag{2.24}$$

onde $F_0(x)$ é o modelo inicial, M é o número total de iterações, $h_m(x)$ é a nova árvore ajustada aos resíduos, e γ_m é a taxa de aprendizado que controla a contribuição de cada árvore.

- **Bayes Ingênuo (em inglês, *Naive Bayes* (NB))**: segundo (NATEKIN; KNOLL, 2013), Bayes ingênuo calcula a probabilidade de cada classe usando o teorema de Bayes, assumindo independência entre os atributos. É eficiente para grandes conjuntos de dados e útil em problemas com muitas classes.

  O teorema de Bayes é expresso como:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{2.25}$$

  onde $P(C_k|x)$ é a probabilidade da amostra x pertencer à classe C_k, $P(x|C_k)$ é a probabilidade de observar x dado que a classe é C_k, $P(C_k)$ é a probabilidade a priori da classe C_k, e $P(x)$ é a probabilidade total de observar x.

  O Bayes Ingênuo assume que os atributos são independentes, simplificando o cálculo de $P(x|C_k)$ como:

$$P(x|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \tag{2.26}$$

  onde x_i é o valor do i-ésimo atributo e n é o número total de atributos.

  A classe final atribuída à amostra x é aquela que maximiza $P(C_k|x)$:

$$\hat{y} = \arg\max_{C_k} P(C_k|x) \tag{2.27}$$

- **Impulsão de Gradiente Extremo (em inglês, *Extreme Gradient Boosting* (XGBoost))**: é uma implementação otimizada do *gradient boosting* que melhora a velocidade e a eficiência do modelo. O XGBoost utiliza técnicas avançadas, como paralelismo durante o treinamento e regularização, para reduzir o *overfitting* e aumentar a precisão da classificação.
  O modelo ajusta os resíduos iterativamente, similar ao *gradient boosting*, mas inclui um termo de regularização em sua função de custo para controlar a complexidade do modelo. A função de custo em XGBoost é definida como:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.28}$$

  onde $l(y_i, \hat{y}_i)$ é a função de perda que mede a diferença entre os valores reais y_i e as previsões $\hat{y}_i$, e $\Omega(f_k)$ é o termo de regularização para a k-ésima árvore, definido como:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.29}$$

  em que T é o número de folhas da árvore, w_j são os pesos das folhas, γ controla a penalização do número de folhas, e λ controla a penalização dos pesos das folhas, ajudando a reduzir o *overfitting*.
  O XGBoost aprimora o treinamento ao calcular uma aproximação de segunda ordem do gradiente para ajustar os modelos:

$$L(\phi) \approx \sum_{i=1}^{n} \left[g_i f(x_i) + \frac{1}{2} h_i f(x_i)^2 \right] + \Omega(f) \tag{2.30}$$

  onde g_i e h_i são o gradiente e o hessiano da função de perda, respectivamente. Essa aproximação de segunda ordem permite um ajuste mais preciso e eficiente (TARWIDI *et al.*, 2023).

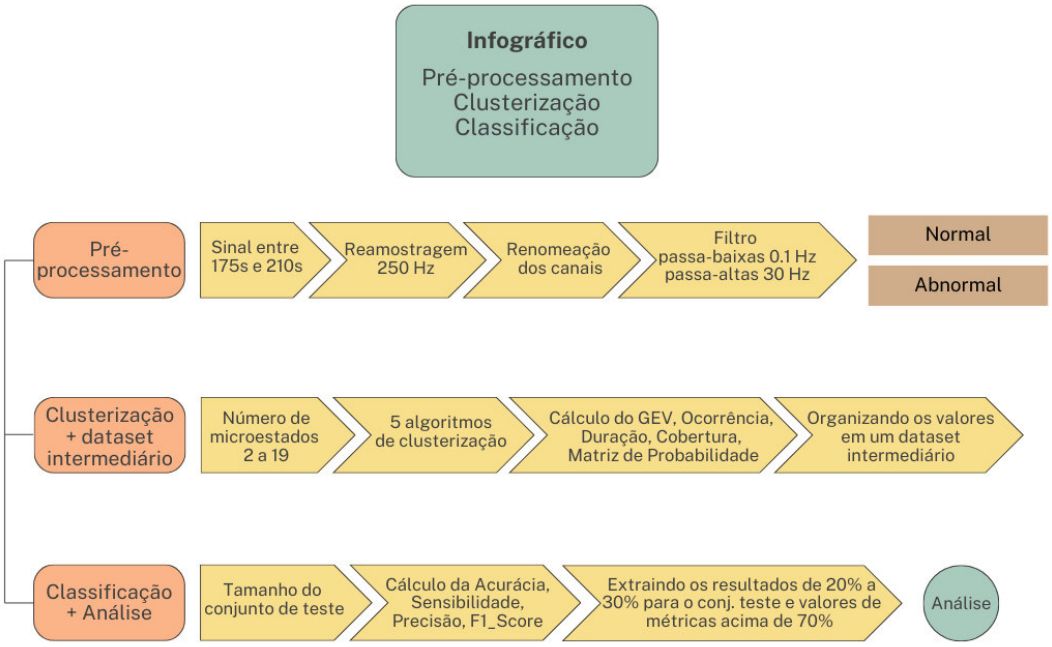
Em resumo, algoritmos de clusterização agrupam dados em subconjuntos homogêneos sem rótulos pré-definidos, identificando padrões ocultos ou não facilmente visíveis, com base em similaridades particulares a cada grupo. Já os algoritmos de classificação, utilizam dados rotulados para treinar modelos que analisam as características e se tornam progressivamente especialistas em relacionar novas amostras a uma categoria pré-estabelecida. Enquanto a clusterização é útil para análise exploratória e descoberta de estruturas nos dados, a classificação é importante para tarefas onde a atribuição de classes é necessária.

3 METODOLOGIA

Esta seção descreve os procedimentos experimentais, incluindo o pré-processamento, a aplicação dos algoritmos e as análises de validação, como mostrado resumidamente na Figura 11. O conjunto de dados *Temple University Hospital Abnormal EEG Corpus* (TUAB), escolhido por sua diversidade com gravações de EEG de indivíduos normais e anormais, foi pré-processado através da remoção de artefatos, aplicação de filtros e normalização dos sinais. Antes da análise principal, realizou-se uma inspeção visual na montagem bipolar longitudinal ("double-banana"), análise das frequências cerebrais e comparação entre sinais filtrados e não filtrados.

Na etapa relativa ao uso das ferramentas de aprendizagem de máquina, foram aplicados algoritmos de clusterização (*K-means*, *K-medoids*, *Modified K-means*, *AAHC*, *DBSCAN*), variando o número de microestados de 2 a 19, para gerar um novo conjunto de dados contendo métricas como Variância Explicada Global (GEV), Ocorrência, Duração, Cobertura e Matriz de Probabilidade de Transição. As clusterizações foram avaliadas por meio das métricas *Silhouette*, Calinski-Harabasz, Davies-Bouldin e Dunn. Em seguida, este conjunto de dados foi utilizado na etapa de classificação, empregando algoritmos como *Random Forest*, *SVM* e *Logistic Regression*, para avaliar a acurácia e outras métricas em conjuntos de teste variando de 10% a 90%. As classificações foram avaliadas utilizando acurácia, sensibilidade (ou revocação), precisão e *F1 score*, garantindo uma análise abrangente da performance dos algoritmos nos dados de EEG.

Figura 11 – Infográfico das etapas realizadas na etapa do fluxo de clusterização e de classificação.



Fonte: Elaborada pelo autor.

3.1 Base de dados

Visando utilizar uma base de dados abrangente e diversa, o conjunto de dados TUAB, descrito por (LóPEZ *et al.*, 2015), foi utilizado. Essa base é um subconjunto da base *TUH EEG Corpus* (TUEG), obtida a partir de *CD-ROMs* contendo as gravações de sessões EEG feitas pelo Departamento de Neurologia do Hospital da Universidade de Temple desde 2002, representando um projeto contínuo de coleta de dados. (OBEID; PICONE, 2016)

3.1.1 Estrutura e Estatística das bases

3.1.1.1 TUH

Os dados de EEG presentes nos CDs foram convertidos para o formato aberto padrão *European Data Format* (EDF). Em seguida, alguns potenciais identificadores foram removidos, incluindo nomes de pacientes e datas de nascimento.

O conjunto extraído foi organizado em uma estrutura de diretórios: a pasta superior que contém 109 pastas numeradas, cada uma contendo subpastas para até 100 pacientes. Cada

subpasta contém sessões de gravação individuais com um ou mais arquivos de dados de EEG (.edf) e o relatório clínico em formato .txt.

Ao total, o corpus da base *TUH EEG* completo possui 16.986 sessões de 10.874 indivíduos únicos, com uma média do número de sessões por paciente de 1,56, e com um único paciente podendo ter até 37 EEGs registrados em um período de 8 meses (Tabela 2). As idades dos pacientes variam de menos de 1 ano a mais de 90 anos, com uma média de 51,6 anos (Tabela 3).

Tabela 2 – Características das sessões.

Característica	**Valor**
Número de indivíduos únicos	10,874
Número total de sessões	16,986
Média de sessões por paciente	1,56
Máximo de sessões por paciente	37

Fonte: o autor, baseado em Obeid e Picone (2016).

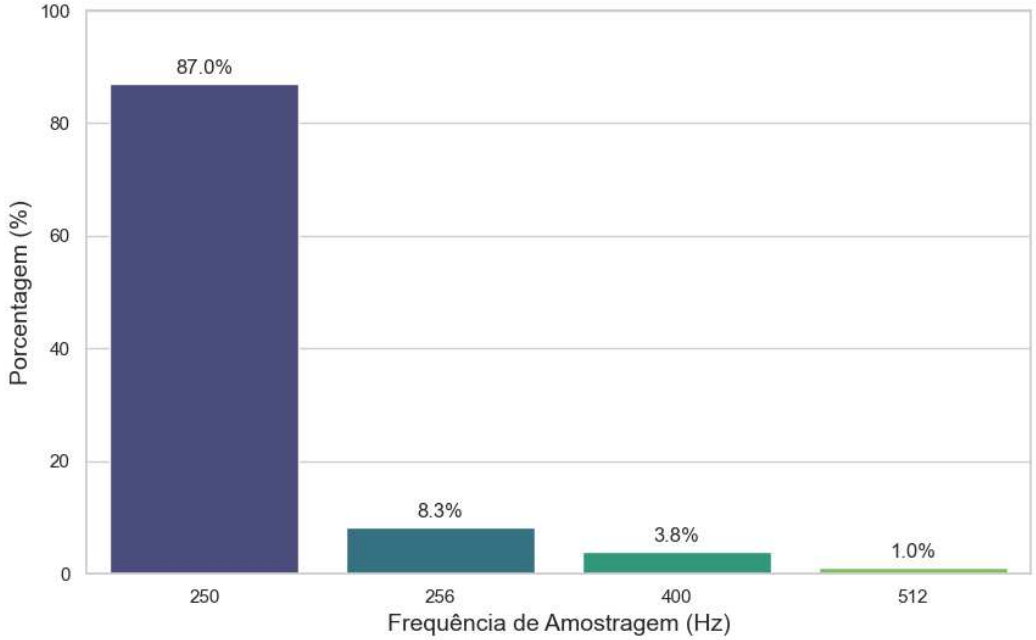
Tabela 3 – Estatísticas da faixa etária dos pacientes.

Estatística	**Valor**
Faixa etária dos pacientes	Menos de 1 ano a mais de 90 anos
Idade média dos pacientes	51,6 anos
Desvio padrão da idade	55,9 anos

Fonte: o autor, baseado em Obeid e Picone (2016).

Os arquivos EDF no *corpus* contêm canais específicos de EEG, além de canais suplementares, EKG (referente à atividade muscular cardíaca), EMG (referente à atividade muscular de repouso e contração) e EOG (referente à atividade da região ocular). O número mais comum de canais de EEG por arquivo EDF é 31, embora haja casos com até 20 canais. A maioria dos dados de EEG foi amostrada a 250 Hz (87%), com o restante sendo amostrado a 256 Hz (8,3%), 400 Hz (3,8%) e 512 Hz (1%), como mostrado na Figura 12.

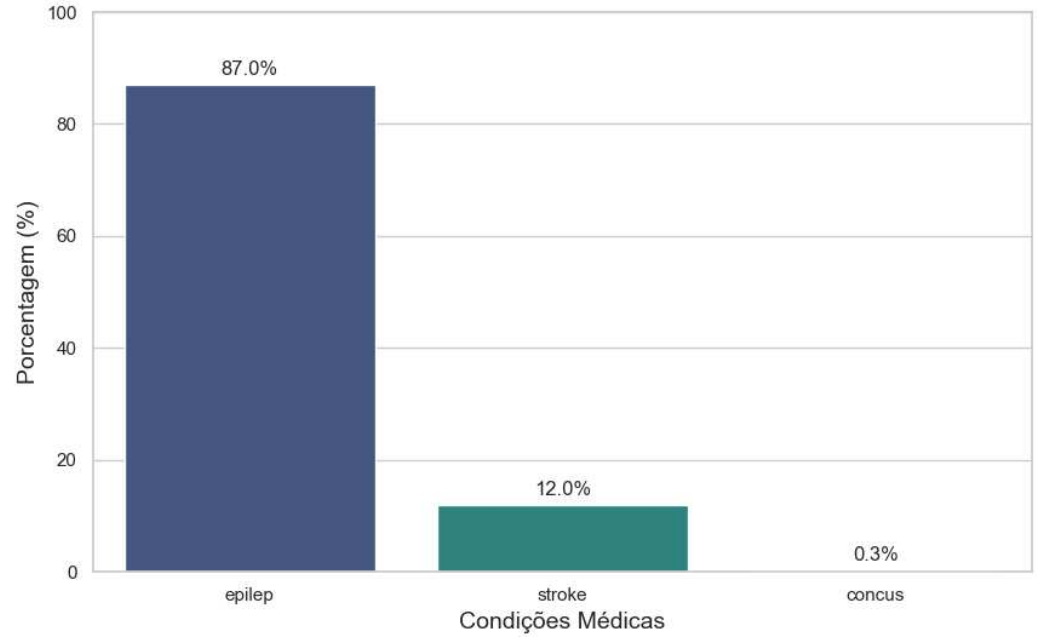
Figura 12 – Distribuição da frequência de amostragem das gravações.



Fonte: Elaborada pelo autor.

As análises iniciais dos relatórios clínicos mostraram a existência de diversas medicações e condições médicas. Em suma, aproximadamente 87% dos relatórios incluíam o texto "*epilep*", referente à epilepsia, enquanto cerca de 12% mencionavam "*stroke*", referente ao Acidente Vascular Cerebral (AVC) e apenas 48 relatórios incluíam "*concus*", referente a casos de concussão, como apresentado na Figura 13.

Figura 13 – Distribuição da frequência das condições clínicas.



Fonte: Elaborada pelo autor.

3.1.1.2 TUAB

A atribuição das classes no TUAB foi feita de forma automática, empregando alguns algoritmos de aprendizagem de máquina aplicados apenas aos sinais EEG (LóPEZ *et al.*, 2015), e consolidada a partir de uma revisão das informações apresentadas nos relatórios médicos, havendo concordância em cerca de 97-100% sobre as classificações das gravações EEG como patológicas ou não-patológicas.

Ainda, a seleção das gravações foi baseada em gênero e idade para obter um conjunto de dados demograficamente equilibrado. Por sessão, apenas as que tinham duração acima de 15 minutos foram consideradas, e, para os pacientes patológicos, as gravações selecionadas foram as que também possuíam atividades mecânicas nos relatórios clínicos.

Em suma, o TUAB compreende 2.993 gravações de 2.329 pacientes (52,09% do sexo feminino), com idade média de 48,55 ± 17,86 anos. O conjunto é quase equilibrado em termos do número de gravações não-patológicas (50,82%) e patológicas (49,18%). O conjunto de treinamento consiste em 2.717 gravações de 2.130 pacientes, enquanto o conjunto de avaliação contém 276 gravações de 251 pacientes, como mostrado nas Tabelas 4 e 5. (KIESSNER *et al.*, 2023)

Tabela 4 – Características gerais.

Característica	**Valor**
Número de indivíduos únicos	2,329
Número total de gravações	2,993
Média de gravações por paciente	1,29
Máximo de gravações por paciente	37

Fonte: o autor, baseado em (KIESSNER *et al.*, 2023)

Tabela 5 – Número de gravações e pacientes no conjunto de dados TUAB.

Conjunto	**Gravações Não-Patológicas**	**Gravações Patológicas**	**Total de Gravações**
Treinamento	1371	1346	2717
Avaliação	150	126	276
Total	**1521**	**1472**	**2993**

Fonte: Adaptado de (KIESSNER *et al.*, 2023)

3.1.2 Disponibilidade do conjunto de dados

Tanto o *TUH* quando o *TUAB* estão disponíveis na internet acessando o *site* https://isip.piconepress.com/projects/nedc/html/tuh_eeg/ do professor responsável pela produção,

organização e melhorias das respectivas bases, mas também é possível acessá-los junto a diversas ferramentas usando os códigos disponíveis em repositórios de código como o *github*, por exemplo, a biblioteca *PyHealth* (https://github.com/sunlabuiuc/PyHealth), entre outros.

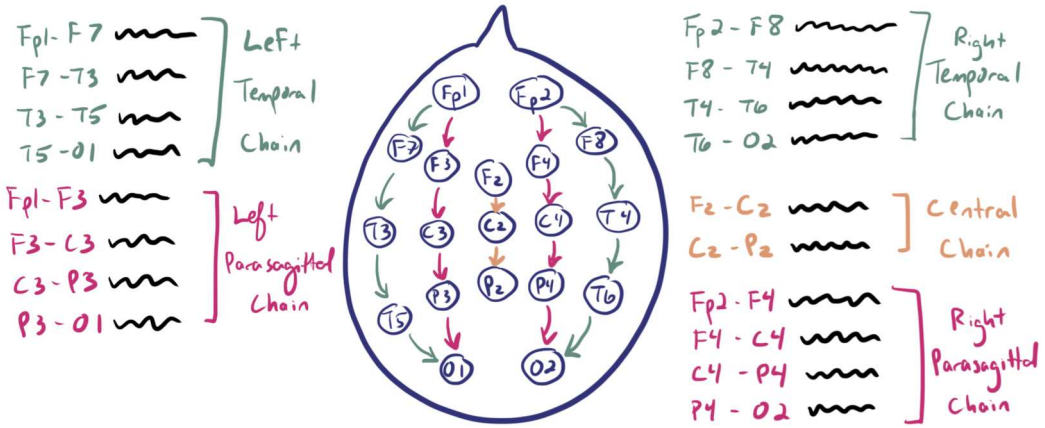
3.2 Análise inicial

Nesta seção, são apresentadas as etapas iniciais de análise dos dados de EEG, focando em três abordagens principais: a análise do formato *double-banana*, a análise das frequências de bandas cerebrais e a comparação entre sinais filtrados e não-filtrados. Essas etapas são fundamentais para uma compreensão detalhada dos dados e para garantir que o pré-processamento e a segmentação dos microestados sejam realizados de maneira eficaz. A seguir, são descritos os métodos e os resultados obtidos em cada uma dessas abordagens.

3.2.1 Análise do Formato Double-Banana

Para a análise do formato *double-banana*, foi realizada uma inspeção visual da montagem bipolar longitudinal. Este formato é utilizado para destacar diferenças de potencial elétrico entre pares de eletrodos, permitindo uma visualização clara das variações de atividade cerebral ao longo do couro cabeludo. A diferença de potencial é calculada entre pares de eletrodos adjacentes, conforme ilustrado na Figura 14.

Figura 14 – Montagem bipolar longitudinal (*double-banana*).



Fonte: Obtido em www.learningeeg.com.

Na Figura 14, os pares de eletrodos são mostrados conforme as diferentes cadeias temporais, parasagitais e centrais, permitindo uma análise detalhada da atividade elétrica em diversas regiões do cérebro.

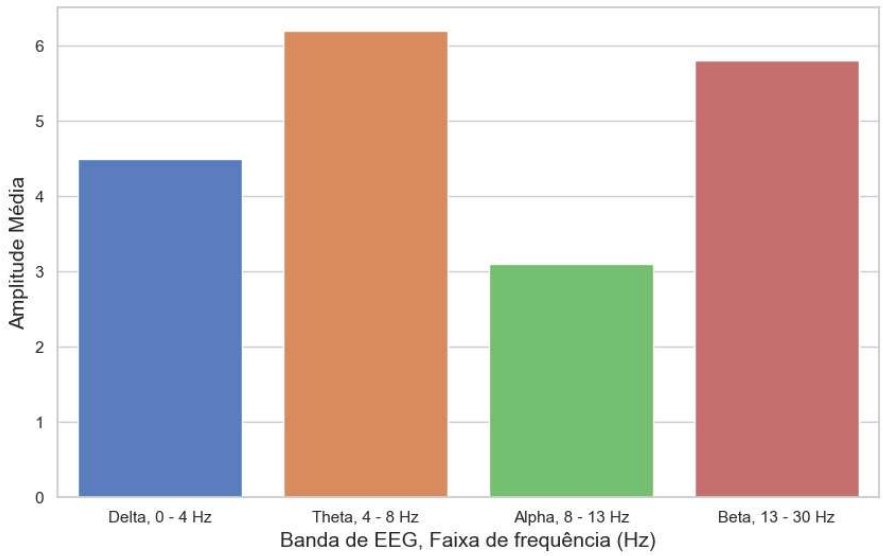
A análise visual das imagens possibilita a identificação de padrões específicos de atividade elétrica associados a diferentes estados mentais e condições neurológicas. Por exemplo, variações anormais na amplitude ou frequência das ondas cerebrais podem ser indicativas de epilepsia, distúrbios do sono, ou outras condições neurológicas. Além disso, a montagem *double-banana* permite observar a sincronização e a coerência entre diferentes regiões do cérebro, oferecendo percepções sobre a conectividade funcional.

3.2.2 Análise das Frequências de Bandas Cerebrais

De acordo com (JOSHI *et al.*, 2019), para analisar as frequências de bandas cerebrais, foi realizada a Transformada Rápida de Fourier (*Fast Fourier Transform*, FFT). Com isso, calculou-se as amplitudes médias das bandas de EEG e geram-se os respectivos gráficos de coluna.

O procedimento foi aplicado à primeira gravação de cada classe (normal e abnormal) e também a um sinal médio de cada classe. A Figura 15 apresenta um exemplo do gráfico gerado.

Figura 15 – Análise das Frequências de Bandas Cerebrais.



Fonte: Elaborada pelo autor.

A análise das amplitudes médias dessas bandas permite inferir sobre o estado de vigília, relaxamento, concentração e possíveis distúrbios neurológicos. Por exemplo, uma maior amplitude nas bandas *theta* e *delta* em um indivíduo acordado pode sugerir disfunções neurológicas ou distúrbios de atenção. Da mesma forma, um aumento significativo na banda

beta pode ser um indicador de estresse ou ansiedade.

Ao comparar as gravações de indivíduos normais e anormais, pode-se observar diferenças significativas nas amplitudes médias das bandas, auxiliando os profissionais de saúde no diagnóstico de possíveis doenças ou anormalidades.

3.3 Aplicação dos algoritmos de aprendizagem de máquina

O objetivo do experimento principal foi analisar e comparar gravações de EEG das classes normal e anormal utilizando técnicas de clusterização de microestados e modelos de classificação. O processo envolveu a extração das gravações, aplicação de etapas de pré-processamento, segmentação dos microestados e avaliação do desempenho dos modelos de classificação com base em métricas. As etapas detalhadas incluem a extração dos dados dos arquivos *.edf*, aplicação de filtros, renomeação de canais, e utilização de diversos algoritmos de clusterização e modelos de classificação para análise das métricas. As etapas envolvidas nesse experimento principal serão discutidas a seguir.

3.3.1 Pré-processamento e execução do fluxo de clusterização e de classificação

Primeiramente, as gravações de EEG foram extraídas dos arquivos *.edf*, considerando apenas o sinal correspondente aos segundos 175 e 210 de cada registro, seguidas por uma reamostragem (*resample*) para uma frequência de 250 Hz. Em seguida, os canais foram renomeados de acordo com o mapeamento apresentado na Tabela 6. Todas essas etapas de pré-processamento foram feitas usando a biblioteca MNE da linguagem de programação *Python*.

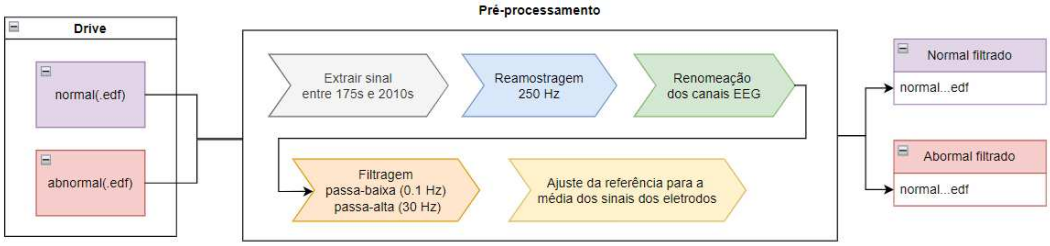
Tabela 6 – Correspondência dos Nomes dos Canais de EEG.

Canal Original	Canal Renomeado
EEG FP1-REF	FP1
EEG FP2-REF	FP2
EEG F3-REF	F3
EEG F4-REF	F4
EEG C3-REF	C3
EEG C4-REF	C4
EEG P3-REF	P3
EEG P4-REF	P4
EEG O1-REF	O1
EEG O2-REF	O2
EEG F7-REF	F7
EEG F8-REF	F8
EEG T3-REF	T3
EEG T4-REF	T4
EEG T5-REF	T5
EEG T6-REF	T6
EEG A1-REF	A1
EEG A2-REF	A2
EEG FZ-REF	FZ
EEG CZ-REF	CZ
EEG PZ-REF	PZ

Fonte: Elaborado pelo autor.

Posteriormente, foi aplicado um filtro passa-faixa com frequência de corte de 0.1 Hz e 30 Hz, respectivamente. Finalmente, foi utilizada a referência média (*average reference*) para todos os sinais. Esse pré-processamento está visualmente resumido na Figura 16.

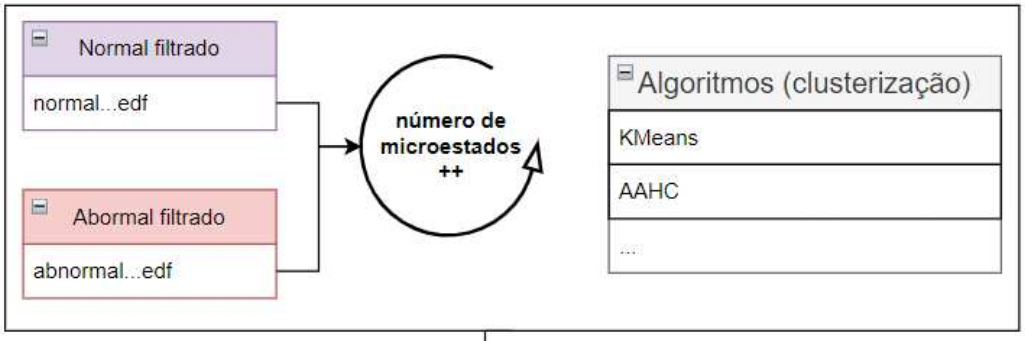
Figura 16 – Pré-processamento.



Fonte: Elaborado pelo autor.

Em seguida, todas as gravações de cada classe foram reunidas em uma respectiva lista e, com isso, foi feita uma sequência de iterações, variando o número de microestados, gerando um conjunto de dados contendo colunas com as métricas GEV, Ocorrência, Duração, Cobertura e Matriz de Probabilidade de Transição, utilizando diferentes algoritmos de clusterização: *K-Means*, *AAHC*, *DBSCAN*, *Kmedoids* e *Modified K-Means*. Essa outra etapa do fluxo está apresentada na Figura 17, onde as listas geradas da etapa de pré-processamento são usadas para gerar um *dataset* intermediário que será a entrada de dados para a etapa de classificação.

Figura 17 – Clusterização e Geração do *dataset* intermediário.

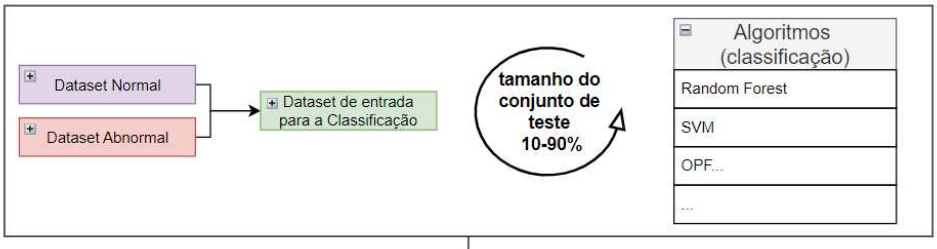


Número de Microestados	Algoritmo	GEV	Ocorrência	Duração	Cobertura	Matriz de Transição de Probabilidade	Rótulo da classe
2	KMeans	Valor	Valor	Valor	Valor	Valor	0
3	KMeans	Valor	Valor	Valor	Valor	Valor	1
...	...	...	...	...	...	...	...

Fonte: Elaborado pelo autor.

Após isso, os novos conjuntos de dados foram concatenados e aplicados a diversos modelos de classificação para analisar as métricas de desempenho. Exemplos de algoritmos de classificação utilizados incluem *Random Forest*, *SVM* e *Logistic Regression*. A Figura 18 resume o fluxo explanado e a Tabela 7 apresenta todos os algoritmos de classificação utilizados.

Figura 18 – Clusterização e Geração do *dataset* intermediário.



Modelo	Número de Microestados	Tamanho do conjunto de teste	Algoritmo de clusterização	Acurácia	Precisão	Sensibilidade	F1_Score	Matriz de Confusão
Random Forest	2	10%	KMeans	Valor	Valor	Valor	Valor	Valor
SVM	19	30%	KMedoids	Valor	Valor	Valor	Valor	Valor
...	...	...	...	...	...	...	...	...

Fonte: Elaborado pelo autor.

Tabela 7 – Algoritmos de Classificação Utilizados.

Algoritmo	**Descrição**
Random Forest	Floresta de Árvores Aleatórias
SVM	Máquina de Vetores de Suporte
Logistic Regression	Regressão Logística
K-Neighbors	K-Vizinhos Mais Próximos
Decision Tree	Árvore de Decisão
Gradient Boosting	Aprimoramento Gradiente
Naive Bayes	Classificador de Bayes Ingênuo
XGBoost	Aumento de Gradiente Extremo
OPF	Floresta de Caminhos Ótimos

Fonte: Elaborada pelo autor.

Além disso, diferentes parâmetros de distância foram utilizados na instanciação do algoritmo OPF em cada iteração, conforme apresentado nas Tabelas 8 e 9, em que x e y representam os respectivos vetores de características de cada amostra. O algoritmo e as distâncias estão descritos e/ou referenciados no *link* do *Github* do *OPFython* (<https://github.com/gugarosa/opfython>). Por fim, os *dataframes* das métricas gerados foram salvos em arquivos *.xlsx* para posterior análise.

Tabela 8 – Parâmetros de Distância Utilizados no Algoritmo OPF - Parte 1.

Parâmetro de Distância	**Fórmula**	**Descrição**
additive_symmetric	$2\sum \frac{(x-y)^2 \cdot (x+y)}{x \cdot y}$	Distância Simétrica Aditiva
average_euclidean	$\sqrt{\frac{\sum (x-y)^2}{N}}$	Distância Euclidiana Média
bhattacharyya	$-\log\left(\sum \sqrt{x \cdot y}\right)$	Distância Bhattacharyya
bray_curtis	$\frac{\sum \lvert x-y \rvert}{\sum (x+y)}$	Distância Bray-Curtis
canberra	$\sum \frac{\lvert x-y \rvert}{\lvert x \rvert + \lvert y \rvert}$	Distância Canberra
chebyshev	$\max(\lvert x_i - y_i \rvert)$	Distância Chebyshev
chi_squared	$\frac{1}{2}\sum \frac{(x-y)^2}{x+y}$	Distância Qui-Quadrado
chord	$\sqrt{2 - 2\frac{\sum (x \cdot y)}{\sqrt{\sum x^2} \cdot \sqrt{\sum y^2}}}$	Distância Chord
clark	$\sqrt{\sum \left(\frac{x-y}{x+y}\right)^2}$	Distância Clark
cosine	$1 - \frac{\sum (x \cdot y)}{\sqrt{\sum x^2} \cdot \sqrt{\sum y^2}}$	Distância Cosseno
dice	$1 - \frac{2\sum (x \cdot y)}{\sum x^2 + \sum y^2}$	Distância Dice
divergence	$2\sum \frac{(x-y)^2}{(x+y)^2}$	Distância Divergente
euclidean	$\sqrt{\sum (x-y)^2}$	Distância Euclidiana
gaussian	$e^{-\gamma\sqrt{\sum (x-y)^2}}$	Distância Gaussiana
gower	$\frac{\sum \lvert x-y \rvert}{N}$	Distância Gower
hamming	$\sum (x_i \neq y_i)$	Distância Hamming
hassanat	$\sum \left(1 - \frac{1+\min(x,y)}{1+\max(x,y)}\right)$	Distância Hassanat
hellinger	$\sqrt{2\sum \left(\sqrt{x} - \sqrt{y}\right)^2}$	Distância Hellinger
jaccard	$\frac{\sum (x-y)^2}{\sum x^2 + \sum y^2 - \sum (x \cdot y)}$	Distância Jaccard
jeffreys	$\sum (x-y)\log\left(\frac{x}{y}\right)$	Distância Jeffreys
jensen	$0.5\sum \left[\frac{x\log(x) + y\log(y)}{2} - \frac{x+y}{2}\log\left(\frac{x+y}{2}\right)\right]$	Distância Jensen
jensen_shannon	$0.5\left[\sum x \log\left(\frac{2x}{x+y}\right) + \sum y \log\left(\frac{2y}{x+y}\right)\right]$	Distância Jensen-Shannon
k_divergence	$\sum x \log\left(\frac{2x}{x+y}\right)$	Distância K-Divergente
kulczynski	$\frac{\sum \lvert x-y \rvert}{\sum \min(x,y)}$	Distância Kulczynski
kullback_leibler	$10^5 \cdot \log(\text{dist} + 1)$	Distância Kullback-Leibler
log_euclidean	$10^5 \cdot \log(\sqrt{\sum (x-y)^2} + 1)$	Distância Log-Euclidiana
log_squared_euclidean	$10^5 \cdot \log\left(\sum (x-y)^2 + 1\right)$	Distância Euclidiana ao Quadrado Logarítmica
lorentzian	$\sum \log(1 + \lvert x-y \rvert)$	Distância Lorentziana
manhattan	$\sum \lvert x-y \rvert$	Distância Manhattan
matusita	$\sqrt{\sum \left(\sqrt{x} - \sqrt{y}\right)^2}$	Distância Matusita
max_symmetric	$\max\left(\sum \frac{(x-y)^2}{x}, \sum \frac{(x-y)^2}{y}\right)$	Distância Simétrica Máxima
mean_censored_euclidean	$\sqrt{\frac{\sum (x-y)^2}{\text{diff}}}$	Distância Euclidiana Média Censurada
min_symmetric	$\min\left(\sum \frac{(x-y)^2}{x}, \sum \frac{(x-y)^2}{y}\right)$	Distância Simétrica Mínima
neyman	$\sum \frac{(x-y)^2}{x}$	Distância Neyman
non_intersection	$0.5\sum \lvert x-y \rvert$	Distância Não Intersecção
pearson	$\sum \frac{(x-y)^2}{y}$	Distância Pearson
sangvi	$2\sum \frac{(x-y)^2}{x+y}$	Distância Sangvi
soergel	$\frac{\sum \lvert x-y \rvert}{\sum \max(x,y)}$	Distância Soergel
squared	$\sum \frac{(x-y)^2}{x+y}$	Distância ao Quadrado
squared_chord	$\sum \left(\sqrt{x} - \sqrt{y}\right)^2$	Distância Chord ao Quadrado
squared_euclidean	$\sum (x-y)^2$	Distância Euclidiana ao Quadrado

Fonte: Elaborada pelo autor, baseando-se na documentação <https://github.com/gugarosa/opfython/blob/master/opfython/math/distance.py>.

Tabela 9 – Parâmetros de Distância Utilizados no Algoritmo OPF - Parte 2.

Parâmetro de Distância	**Fórmula**	**Descrição**
statistic	$\sum \frac{x-\frac{x+y}{2}}{\frac{x+y}{2}}$	Distância Estatística
topsoe	$\sum x \log\left(\frac{2x}{x+y}\right) + \sum y \log\left(\frac{2y}{x+y}\right)$	Distância Topsoe
vicis_symmetric1	$\sum \frac{(x-y)^2}{\min(x,y)^2}$	Distância Vicis Simétrica 1
vicis_symmetric2	$\sum \frac{(x-y)^2}{\min(x,y)}$	Distância Vicis Simétrica 2
vicis_symmetric3	$\sum \frac{(x-y)^2}{\max(x,y)}$	Distância Vicis Simétrica 3
vicis_wave_hedges	$\sum \frac{\lvert x-y\rvert}{\min(x,y)}$	Distância Vicis-Wave-Hedges

Fonte: Elaborada pelo autor, baseando-se na documentação <https://github.com/gugarosa/opfython/blob/master/opfython/math/distance.py>.

As métricas de desempenho analisadas incluem acurácia, precisão, sensibilidade (*recall*) e *F1 score*, conforme descrito na Tabela 10.

Tabela 10 – Métricas de Desempenho Utilizadas na Classificação.

Métrica	**Descrição**
Acurácia	Proporção de verdadeiros resultados (positivos e negativos)
Precisão	Proporção de verdadeiros positivos entre os resultados positivos
Sensibilidade	Capacidade de identificar todos os verdadeiros positivos
F1 Score	Média harmônica da precisão e da sensibilidade

Fonte: Elaborada pelo autor.

3.3.2 Análise restrita

Para a continuação do experimento principal, dos arquivos *.xlsx* gerados para cada algoritmo, foram selecionados aqueles que obtiveram valores de métricas de desempenho (acurácia, precisão, sensibilidade e *F1 score*) acima de 70%. A seguir, para cada algoritmo e microestado dos dados extraídos dos arquivos, foi realizada a re-segmentação das gravações de EEG das respectivas classes (normal e *abnormal*). A partir dessa nova segmentação, foram calculadas as seguintes métricas: *Silhouette Score*, *Davies-Bouldin Index*, *Calinski-Harabasz Index* e *Dunn Index*.

Figura 19 – Fluxo da segunda parte da etapa principal do experimento.

Fonte: Elaborado pelo autor.

Esses procedimentos permitiram uma análise detalhada e robusta dos microestados, levando em consideração diferentes abordagens de clusterização e modelos de classificação, assegurando a validade e a confiabilidade dos resultados obtidos.

3.4 Métricas

Nesta seção, serão explicadas as métricas usadas na avaliação das etapas da parte principal do experimento, além das envolvidas na etapa intermediária entre a clusterização e a classificação, as quais servirão como *features* dos conjuntos de treino e teste submetidos aos algoritmos de classificação.

3.4.1 Clusterização

Para a etapa de clusterização, foram utilizadas métricas de análise da qualidade do agrupamento feito pelos algoritmos, sendo obtidas pelo uso de funções do módulo *metrics* da biblioteca *sci-kit learn* da linguagem de programação *python*. Essas métricas analisam a similaridade entre os *clusters* gerados, proporcionando informação quantitativa da coesão entre os elementos e separação *inter-clusters*. As métricas utilizadas são:

3.4.1.1 Silhouette Score

O *Silhouette Score* é uma métrica que avalia a coesão e a separação dos *clusters*. Ela é calculada com base na proximidade média *intra-cluster* e na distância média ao *cluster* mais próximo. A fórmula é dada por:

$$\text{Silhouette} = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3.1}$$

onde $a(i)$ é a distância média entre o ponto i e todos os outros pontos do mesmo *cluster*, e $b(i)$ é a distância média entre o ponto i e todos os pontos do *cluster* mais próximo. Um valor de *Silhouette Score* próximo de 1 indica que os pontos estão bem agrupados dentro dos seus *clusters* e que os *clusters* estão bem separados entre si. (ROUSSEEUW, 1987)

No entanto, essa métrica pode apresentar fraquezas em situações onde os *clusters* têm formas não esféricas ou variam significativamente em tamanho e densidade. Nesses casos, o *Silhouette Score* pode não refletir adequadamente a verdadeira estrutura dos dados, levando a conclusões errôneas sobre a qualidade do agrupamento.

3.4.1.2 Calinski-Harabasz Index

O *Calinski-Harabasz Index* é uma métrica que avalia a qualidade dos *clusters* com base na soma da dispersão *inter-clusters* e na soma da dispersão *intra-cluster*. A fórmula é dada por:

$$\text{CH} = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{N-k}{k-1} \tag{3.2}$$

onde $\text{tr}(B_k)$ é o traço da matriz de dispersão entre *clusters*, $\text{tr}(W_k)$ é o traço da matriz de dispersão *intra-cluster*, N é o número total de pontos, e k é o número de *clusters*. Um valor mais alto do *Calinski-Harabasz Index* indica uma melhor definição dos *clusters*, sugerindo que os *clusters* são densos e bem separados. (LI *et al.*, 2020)

No entanto, essa métrica pode ser sensível ao número de *clusters*, tendendo a favorecer soluções com mais *clusters*, mesmo que não sejam significativas do ponto de vista prático.

3.4.1.3 Davies-Bouldin Index

O *Davies-Bouldin Index* mede a média da razão entre a soma da dispersão dentro de cada *cluster* e a separação entre os *clusters*. A fórmula é dada por:

$$\text{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \tag{3.3}$$

onde s_i é a dispersão média *intra-cluster* para o *cluster i*, d_{ij} é a distância entre os centroides dos *clusters i* e j. Um valor mais baixo do *Davies-Bouldin Index* indica uma melhor separação entre os *clusters*, sugerindo que os *clusters* são compactos e bem separados. (BANDITWATTANAWONG; MASDISORNCHOTE, 2021)

Entretanto, essa métrica pode ser influenciada negativamente por *clusters* de formas e tamanhos muito diferentes, o que pode resultar em uma avaliação inadequada da qualidade do agrupamento.

3.4.1.4 Dunn Index

O *Dunn Index* avalia a compacidade e a separação dos *clusters*, calculando a razão entre a menor distância entre dois *clusters* e o maior diâmetro de um *cluster*. A fórmula é dada por:

$$\text{Dunn} = \frac{\min_{1 \leq i < j \leq k} d(c_i, c_j)}{\max_{1 \leq i \leq k} d(c_i)} \tag{3.4}$$

onde $d(c_i, c_j)$ é a distância entre os centroides dos *clusters i* e j, $d(c_i)$ é o diâmetro do *cluster i* e k é o número de *clusters*. Um valor mais alto do *Dunn Index* indica *clusters* bem separados e compactos, sendo útil para detectar a presença de *clusters* com formas irregulares e assegurar que os *clusters* sejam bem definidos (Ben Ncir *et al.*, 2021).

No entanto, essa métrica pode ser afetada por *outliers* e pode não ser adequada para conjuntos de dados muito grandes devido à sua complexidade computacional.

3.4.2 Etapa intermediária

Em seguida, algumas grandezas foram calculadas para avaliar a qualidade dos *clusters* gerados pelos algoritmos, formando um conjunto de dados para a etapa de classificação. Estas métricas fornecem uma análise detalhada da dinâmica temporal dos microestados de EEG, são elas: Variância Global Explicada (*global explained variance*, GEV), Ocorrência (*occurrence*), Duração (*duration*), Cobertura (*coverage*) e Matriz de Probabilidade de Transição (*transition probability matrix*).

3.4.2.1 Variância Global Explicada (Global Explained Variance, GEV)

A Variância Explicada Global (em inglês, *global explained variance*, GEV) quantifica a proporção da variância total dos dados que é explicada pelos microestados identificados. O cálculo da sequência de microestados é idêntico para todos os métodos de clusterização, seguindo uma abordagem de "o vencedor leva tudo"(PASCUAL-MARQUI *et al.*, 1995). A GEV mede o percentual de variância dos dados explicada por um conjunto de mapas de microestados, fornecendo assim uma métrica para avaliar o quão bem os microestados obtidos correspondem ao conjunto de dados.

A fórmula para o valor da GEV de um mapa de microestado específico com rótulo l é:

$$\mathrm{GEV}_l = \frac{\sum_i \sigma_i^2 C_{il}^2 \delta_{l,L_i}}{\sum_i \sigma_i^2} \tag{3.5}$$

onde: σ_i^2 é a variância no tempo i; C_{il}^2 é a correlação quadrada no tempo i para o microestado l; δ_{l,L_i} é o delta de *Kronecker*, ou seja, $\delta_{l,L_i} = 1$ se $L_i = l$, e $\delta_{l,L_i} = 0$ caso contrário.

A variância explicada global total (GEV) é a soma dos valores de GEV para todos os mapas de microestados:

$$\mathrm{GEV}_{\mathrm{total}} = \sum_l \mathrm{GEV}_l \tag{3.6}$$

Quanto maior o valor dessa métrica, maior a proporção da variância dos dados explicada pelos microestados, indicando que os microestados capturam uma parte significativa da variabilidade presente nos dados de EEG. (WEGNER *et al.*, 2018)

No entanto, essa métrica pode ser enganosa se os dados contiverem variabilidade que não esteja relacionada aos microestados de interesse. Por exemplo, artefatos de movimento ou ruído podem aumentar a variância total, resultando em um valor de GEV mais baixo, mesmo que os microestados identificados sejam altamente correspondentes.

3.4.2.2 Ocorrência (Occurrence)

A métrica Ocorrência (em inglês, *occurrence*) mede a frequência com que cada microestado aparece durante as gravações de EEG. É expressa como a proporção do tempo total

em que um microestado específico é dominante. Esta métrica é útil para identificar padrões comuns de atividade cerebral e comparar a prevalência de diferentes microestados entre grupos de sujeitos. A fórmula é dada por:

$$\frac{\sum_{i=1}^{N} \delta(L_i, l)}{N} \tag{3.7}$$

onde N é o número total de amostras; $\delta(L_i, l)$ é a função delta de *Kronecker* que retorna 1 se o microestado L_i no tempo i for igual a l e 0 caso contrário.

Quanto maior o valor dessa métrica, mais frequentemente o microestado aparece nas gravações de EEG, indicando sua relevância e prevalência na atividade cerebral registrada. (MICHEL; KOENIG, 2018)

No entanto, a alta ocorrência de um microestado não necessariamente indica sua relevância ou qualidade. Por exemplo, se um microestado específico ocorrer com muita frequência devido a artefatos ou estados transitórios irrelevantes, a métrica Ocorrência pode superestimar a importância desse microestado.

3.4.2.3 Duração (Duration)

A Duração (em inglês, *duration*) refere-se ao tempo médio que um microestado permanece ativo antes de transitar para outro microestado (MACKINTOSH *et al.*, 2020). Esta métrica fornece *insights* sobre a estabilidade temporal dos microestados e pode ser utilizada para distinguir entre estados cerebrais de curta e longa duração. A fórmula é dada por:

$$\frac{1}{n_k} \sum_{j=1}^{n_k} d_j \tag{3.8}$$

onde n_k é o número de episódios do microestado k; e d_j é a duração do episódio j.

Quanto maior o valor dessa métrica, maior é a estabilidade do microestado, indicando que ele permanece ativo por períodos mais longos. Isso pode ser interpretado como um sinal de que o microestado é representativo de um estado cerebral estável e consistente.

No entanto, essa métrica pode ser influenciada por episódios longos e anômalos de microestados. Por exemplo, se um microestado for ativado por um longo período devido a um artefato ou condição específica, a métrica Duração pode superestimar a estabilidade desse microestado.

3.4.2.4 Cobertura (Coverage)

A métrica Cobertura (em inglês, *coverage*) refere-se à proporção do tempo total das gravações de EEG em que cada microestado específico está ativo. Esta métrica fornece uma visão quantitativa de quanto tempo cada microestado domina a atividade cerebral ao longo do tempo de gravação. A fórmula para calcular a Cobertura é:

$$\frac{\sum_{i=1}^{N} \delta(L_i, l)}{N} \tag{3.9}$$

onde N é o número total de amostras; $\delta(L_i, l)$ é a função delta de *Kronecker* que retorna 1 se o microestado L_i no tempo i for igual a l e 0 caso contrário.

A cobertura indica a dominância temporal de um microestado particular durante a gravação de EEG. Quanto maior o valor dessa métrica, mais tempo o microestado permanece ativo em relação ao tempo total de gravação, sugerindo sua relevância na atividade cerebral do sujeito. (LASSI *et al.*, 2023)

No entanto, um valor elevado de cobertura não necessariamente indica um microestado de alta qualidade ou importância funcional. Microestados com alta cobertura podem refletir estados cerebrais passivos ou de inatividade, ou ainda podem ser influenciados por artefatos ou outros fatores externos não relacionados à atividade neural desejada.

3.4.2.5 Matriz de Probabilidade de Transição (Transition Probability Matrix)

A métrica Matriz de Probabilidade de Transição (em inglês, *transition probability matrix*) é usada para quantificar a probabilidade de transição de um microestado para outro ao longo do tempo. Essa métrica captura a dinâmica temporal entre diferentes microestados, fornecendo uma representação mais detalhada de como os microestados mudam durante as gravações de EEG.

A matriz de probabilidade de transição é uma matriz quadrada P de dimensão $K \times K$, onde K é o número de microestados. Cada elemento P_{ij} da matriz representa a probabilidade de transição do microestado i para o microestado j (PANDE *et al.*, 2010). A fórmula para calcular cada elemento da matriz é:

$$P_{ij} = \frac{\text{Número de transições de } i \text{ para } j}{\text{Número total de transições de } i} \tag{3.10}$$

onde P_{ij} é a probabilidade de transição do microestado i para o microestado j; o numerador é o número de transições observadas do microestado i para o microestado j; e o denominador é o número total de transições a partir do microestado i.

Essa métrica é útil para entender a sequência de ativação dos microestados e pode revelar padrões de transição específicos que são característicos de diferentes estados cerebrais ou condições patológicas. Por exemplo, padrões anormais de transição entre microestados podem ser indicativos de condições neurológicas específicas.

3.4.3 Classificação

Neste trabalho, para a etapa de classificação, foram utilizadas quatro famosas métricas estatísticas (HASTIE *et al.*, 2009) quando se trata de avaliar os algoritmos de classificação, sendo obtidas pelo uso de funções do módulo *metrics* da biblioteca *sci-kit learn*, são elas: Acurácia, Sensibilidade, Precisão, *F1 Score*. Além disso, em cada execução foi calculada a matriz de confusão com o intuito de facilitar a visualização e entender como os modelos interpretaram os dados no conjunto de testes.

Antes da explicação detalhada de cada métrica, é válido diferenciar conceitos que serão usados para a definição matemática de todas as métricas que ajudam na visualização da performance dos modelos, sendo eles:

- VP (verdadeiros positivos): refere-se à quantidade de amostras que foram corretamente classificadas como pertencentes à classe positiva.
- FN (falsos negativos): refere-se à quantidade de amostras que pertencem à classe positiva, mas foram incorretamente classificadas como pertencentes à classe negativa.
- FP (falsos positivos): refere-se à quantidade de amostras que pertencem à classe negativa, mas foram incorretamente classificadas como pertencentes à classe positiva.
- VN (verdadeiros negativos): refere-se à quantidade de amostras que foram corretamente classificadas como pertencentes à classe negativa.

A matriz de confusão, ilustrada na Tabela 11, é uma ferramenta visual que ajuda a entender esses conceitos. Cada elemento da matriz representa uma contagem de casos em uma das quatro categorias mencionadas acima. A disposição dos elementos na matriz de confusão é a seguinte:

Tabela 11 – Matriz de confusão base.

Verdade	Classificação Positiva	Classificação Negativa
Positiva	VP	FN
Negativa	FP	VN

Fonte: Elaborada pelo autor.

A partir dessa matriz, é possível calcular diversas métricas de performance dos modelos de classificação, como acurácia, precisão, sensibilidade e *F1 score*, que serão detalhadas nas subseções a seguir.

3.4.3.1 Acurácia

A acurácia (*accuracy*) é uma métrica fundamental em problemas de classificação, representando a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões. É uma métrica intuitiva que indica a capacidade geral do modelo em classificar corretamente as amostras. A fórmula para calcular a acurácia é dada por:

$$\frac{VP+VN}{VP+VN+FP+FN} \tag{3.11}$$

Quanto maior o valor da acurácia, melhor é o desempenho do modelo, indicando que ele consegue classificar corretamente uma maior proporção de amostras.

No entanto, a acurácia pode ser uma métrica enganosa em conjuntos de dados desbalanceados, onde uma classe pode ser muito mais frequente que a outra. Nesses casos, um modelo pode apresentar alta acurácia simplesmente por favorecer a classe majoritária, ignorando a classe minoritária. Portanto, é importante considerar outras métricas para uma avaliação mais abrangente do desempenho do modelo.

3.4.3.2 Sensibilidade

A sensibilidade, também conhecida como *recall*, é uma métrica importante em problemas de classificação, especialmente quando a identificação correta dos exemplos positivos é crucial. Ela representa a proporção de verdadeiros positivos (VP) entre o total de exemplos que realmente pertencem à classe positiva. A fórmula para calcular a sensibilidade é dada por:

$$\frac{VP}{VP+FN} \tag{3.12}$$

Quanto maior o valor da sensibilidade, melhor é o desempenho do modelo em identificar corretamente os exemplos positivos. A sensibilidade é particularmente útil em cenários onde a falha em detectar exemplos positivos pode ter consequências graves, como na detecção de doenças.

No entanto, um valor alto de sensibilidade pode ser alcançado às custas de uma maior taxa de falsos positivos, por isso é importante considerar essa métrica juntamente com outras para uma avaliação equilibrada do modelo.

3.4.3.3 Precisão

A precisão (*precision*) é uma métrica que avalia a proporção de verdadeiros positivos (VP) entre todas as amostras que foram classificadas como positivas pelo modelo. Essa métrica é crucial em cenários onde a relevância dos exemplos positivos é alta, como em sistemas de recomendação e detecção de spam. A fórmula para calcular a precisão é dada por:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3.13}$$

Um valor alto de precisão indica que o modelo tem uma baixa taxa de falsos positivos, o que significa que a maioria das previsões positivas são corretas.

No entanto, a precisão deve ser analisada em conjunto com a sensibilidade (*recall*) para fornecer uma visão completa do desempenho do modelo, especialmente em conjuntos de dados desbalanceados.

3.4.3.4 F1 Score

O *F1 score* é uma métrica que combina a precisão (*precision*) e a sensibilidade (*recall*) em uma única medida de desempenho. Ele é especialmente útil em cenários de classificação desbalanceada, onde é importante considerar tanto os falsos positivos quanto os falsos negativos. O *F1 score* é a média harmônica da precisão e da sensibilidade, sendo calculado da seguinte forma:

$$2 \cdot \frac{\text{Precisão} \cdot \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \tag{3.14}$$

O *F1 score* varia de 0 a 1, onde um valor mais próximo de 1 indica um bom equilíbrio entre precisão e sensibilidade. Esta métrica é particularmente útil quando os custos dos falsos

positivos e dos falsos negativos são significativos e devem ser balanceados. A métrica *F1 score* é amplamente utilizada em problemas de classificação como detecção de fraudes, diagnósticos médicos e sistemas de recomendação.

3.5 Recursos computacionais - *hardware* e *software*

Para a realização dos experimentos e análises deste trabalho, utilizou-se o *Google Colab.*, que oferece acesso a recursos computacionais robustos sem custos adicionais. As especificações detalhadas do ambiente de execução são descritas a seguir:

- **CPU**: A instância do *Google Colab.* utilizou uma CPU *Intel Xeon* com 2 *vCPUs* operando a 2.20 *GHz*.
- **Memória RAM**: Foram disponibilizados aproximadamente 13 *GB* de *RAM* para execução dos experimentos.
- ***GPU***: Para aceleração de tarefas de aprendizado de máquina, foi utilizada uma *GPU NVIDIA Tesla K80* com 12 *GB* de *VRAM* (memória de vídeo *RAM GDDR5*).
- **Tempo de Execução**: Cada sessão do *Colab* tem uma duração máxima de 12 horas contínuas. Após esse período, a sessão pode ser interrompida e é necessário reiniciar a instância para continuar o trabalho. Além disso, se a instância permanecer inativa por 90 minutos, ela será desconectada automaticamente.
- **Armazenamento**: A instância oferece aproximadamente 33 *GB* de espaço em disco disponível para armazenamento temporário de arquivos durante as sessões.
- **Versão do *Python***: A versão do *Python* utilizada nas instâncias do *Google Colab.* é a 3.7 ou superior, o que garante compatibilidade com as bibliotecas mais recentes e suporte a funcionalidades modernas da linguagem.

O uso do *Google Colab.* proporcionou uma plataforma eficiente para o desenvolvimento e execução dos modelos de aprendizado de máquina, permitindo o processamento de grandes volumes de dados e a execução de algoritmos complexos com rapidez e eficiência. Essa infraestrutura foi essencial para a realização das análises e experimentos descritos neste trabalho.

4 RESULTADOS

Esta seção traz a exploração detalhada dos resultados observados nos experimentos, inicialmente, fazendo a ánalise qualitativa do conjunto de dados, focando no formato de montagem bipolar longitudinal ('*double-banana*') e nas frequências das ondas cerebrais *delta*, *theta*, *alfa* e *beta*, comparando-os entre as classes normal e *abnormal* do referido TUAB EEG. Em seguida, a etapa principal do experimento, trazendo a análise quantitativa das métricas de desempenho frente aos diferentes números de microestados para os modelos de classificação — acurácia, precisão, entre outras, e para os modelos de clusterização - *sillhouette*, *Dunn*, entre outras.

4.1 Análise qualitativa - montagem bipolar longitudinal (*double-banana*)

A análise clínica da montagem bipolar longitudinal é feita com a visualização da atividade elétrica de regiões específicas do cérebro (parietal, occipital, ...) ao longo do tempo de gravação do EEG de forma simultânea ou posterior à realização do exame.

Profissionais da saúde relacionados ao contexto, por exemplo neurologistas, examinam as forma das ondas, amplitudes e comportamento elétrico local durante um recorte do tempo do exame. Com isso, ele pode comparar com o conhecimento prático e com os padrões determinados pela literatura médica e inferir diagnósticos de anormalidade e de doença da seguinte forma:

- uma vez que são conhecidos os aspectos da série temporal para a condição de ausência de doença (normal), logo, uma anomalia em determinada região pode alertar o médico e incentivar a continuidade da investigação clínica;
- de forma semelhante ao item anterior, uma vez conhecidos os aspectos relacionados a uma doença específica, o profissional pode solicitar outros testes que atestem a enfermidade ou assumí-la conclusivamente.

Além disso, a visualização gráfica da montagem bipolar longitudinal ao longo de várias sessões de EEG permite ao profissional perceber os efeitos de tratamentos farmacológicos ou terapêuticos, caso haja regularidade e controle na realização das gravações de EEG. A exemplificar, a percepção de redução na amplitude de determinados padrões observados em uma determinada região pode indicar se o tratamento está dando resultados positivos, indiferentes ou negativos ao paciente, fornecendo um parecer acerca de possíveis ajustes, personalizando a

experiência terapêutica e possibilitando melhor tratamento das condições neurológicas.

4.1.1 Análise e Visualização

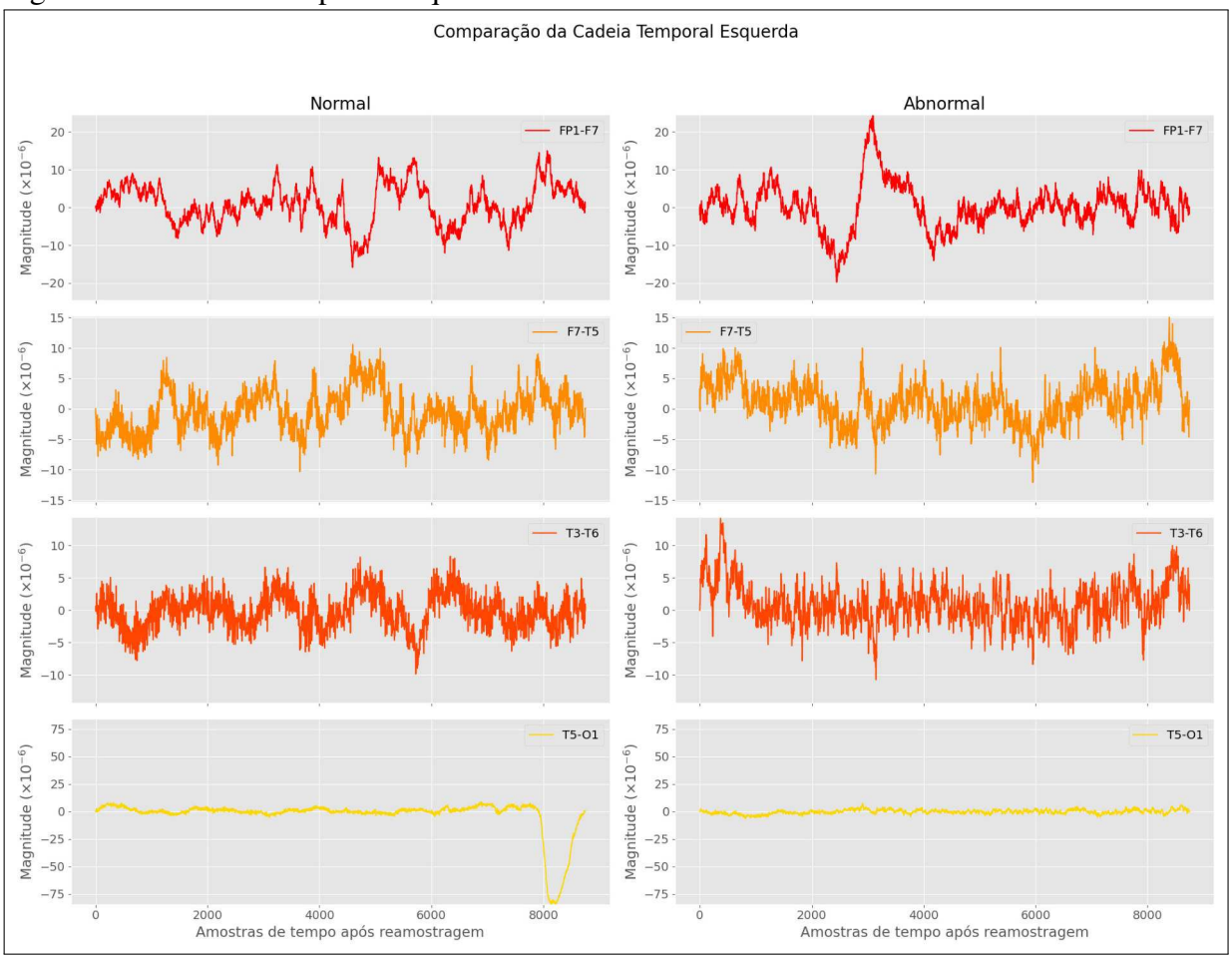
Inicialmente foi feita uma plotagem de 35 segundos, iniciando do segundo 175 até o 210, da parte da primeira gravação EEG para as respectivas classes normal e *abnormal*, porém percebeu-se que essa abordagem enfraqueceria o fator comparativo dessa técnica. Portanto, foi calculada a média do sinal elétrico do trecho escolhido, considerando todas as gravações EEG para cada classe e feita a plotagem comparativa lado a lado entre as regiões. Dessa forma, foi possível produzir os gráficos e as análises dos gráficos resultantes de cada região do *double-banana*, ressaltando que essas plotagens não foram posteriormente averiguadas por um profissional adequado, mas que estão embasadas no que está presente nos trabalhos acadêmicos lidos.

Nas gravações EEG da classe *abnormal*, as anomalias nos gráficos oferecem sugestões de possíveis doenças, por exemplo, picos acentuados e variabilidade aumentada podem indicar atividade epiléptica, se forem recorrentes e focadas em um local; em casos de AVC, pode indicar as regiões onde houve interrupção do fluxo sanguíneo e, portanto, dano neuronal; e, em casos de concussões, podem indicar desordem da atividade cerebral devido ao trauma. Esses aspectos podem ser identificados, por exemplo, nos gráficos da classe *abnormal* nas Figuras 20 (FP1-F7), 21 (C3-P3), 22 (FP2-F8), 23 (C4-P4) e 24 (CZ-PZ).

Nota-se, ainda, que alguns gráficos relacionados à classe normal apresentaram um pico não esperado, por exemplo, como visto nas Figuras 20 (T5-O1), 21 (P3-O1), 22 (T6-O2), 23 (P4-O2) e 24 (FZ-CZ). Tais picos podem ser relacionados a estímulos momentâneos que o paciente teve no momento da gravação.

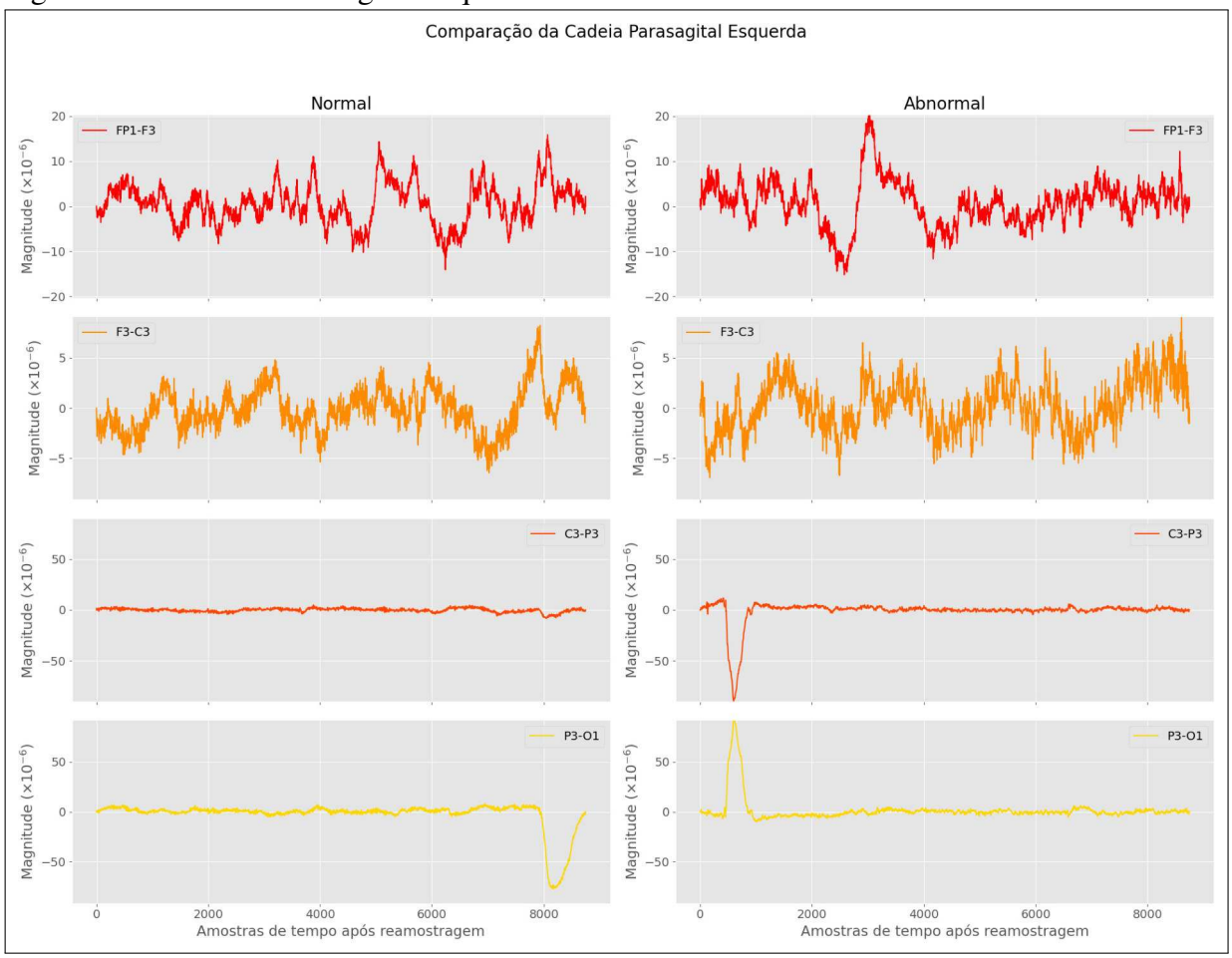
O intuito ao trazer essa análise ao corpo do trabalho é compartilhar formas de analisar conjuntos de dados EEG e, com isso, fornecer instruções e códigos para análise visual das características intrínsecas às gravações de pacientes com distúrbios neurológicos.

Figura 20 – Cadeia Temporal Esquerda.



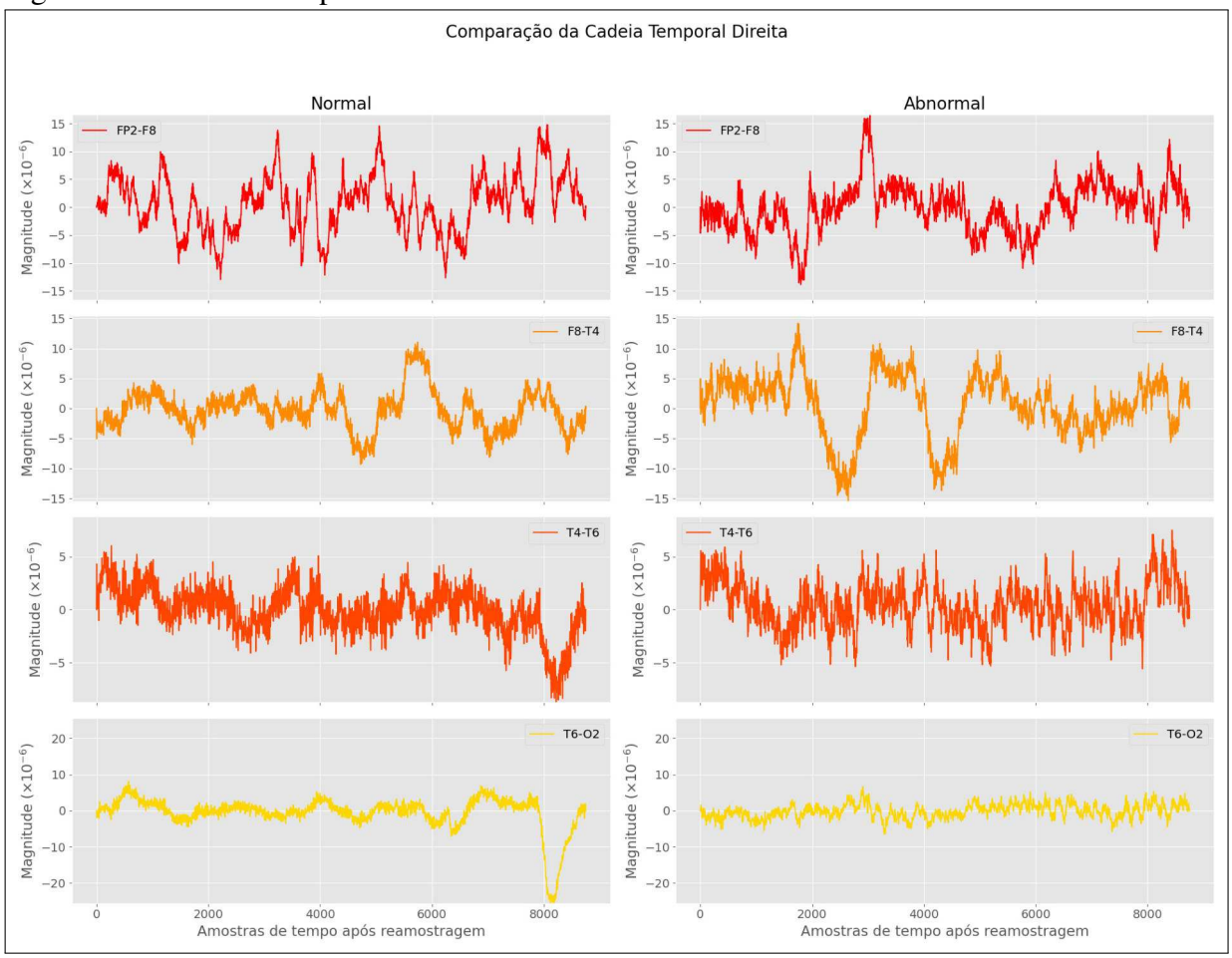
Fonte: elaborado pelo autor (2024).

Figura 21 – Cadeia Parassagital Esquerda.



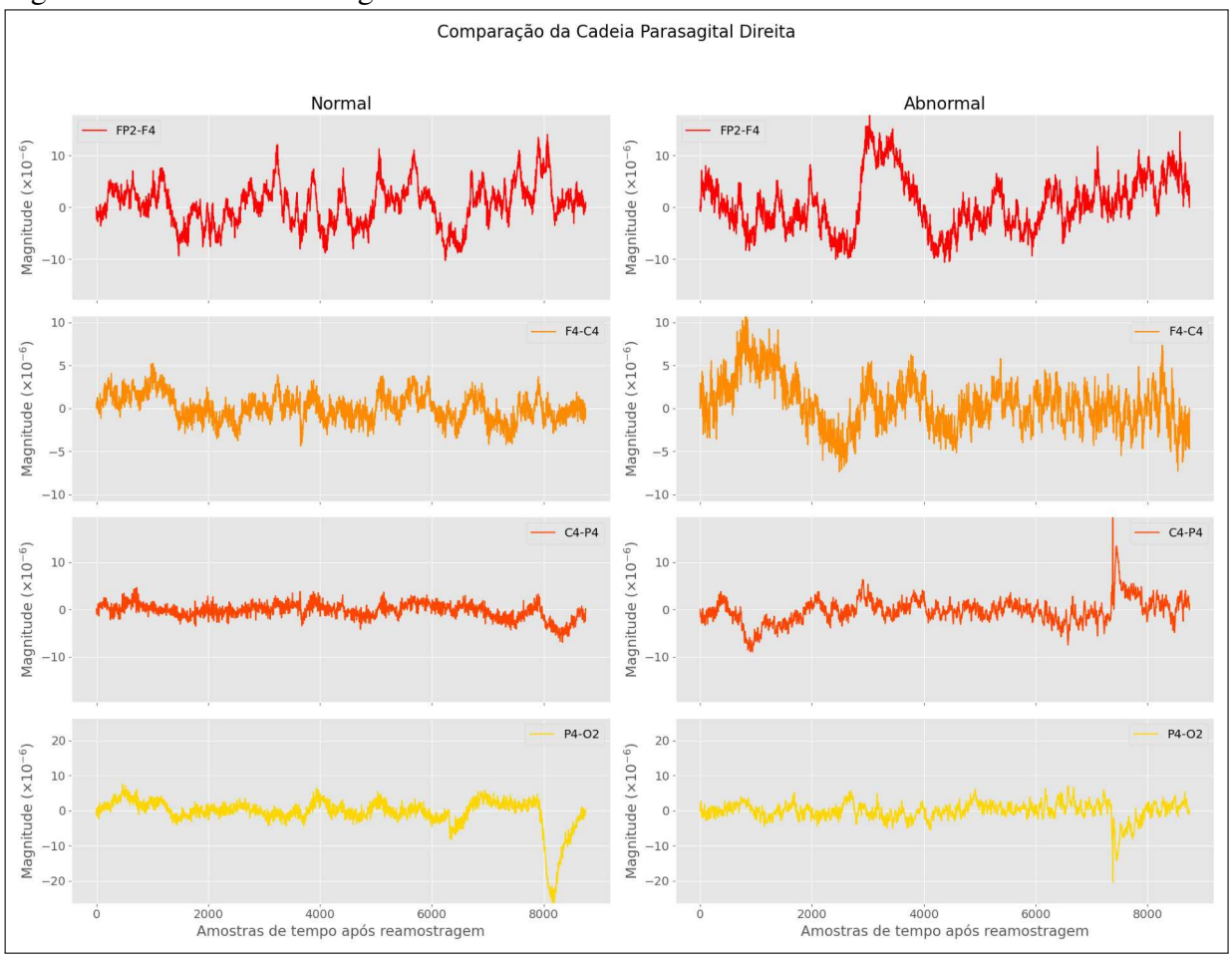
Fonte: elaborado pelo autor (2024).

Figura 22 – Cadeia Temporal Direita.



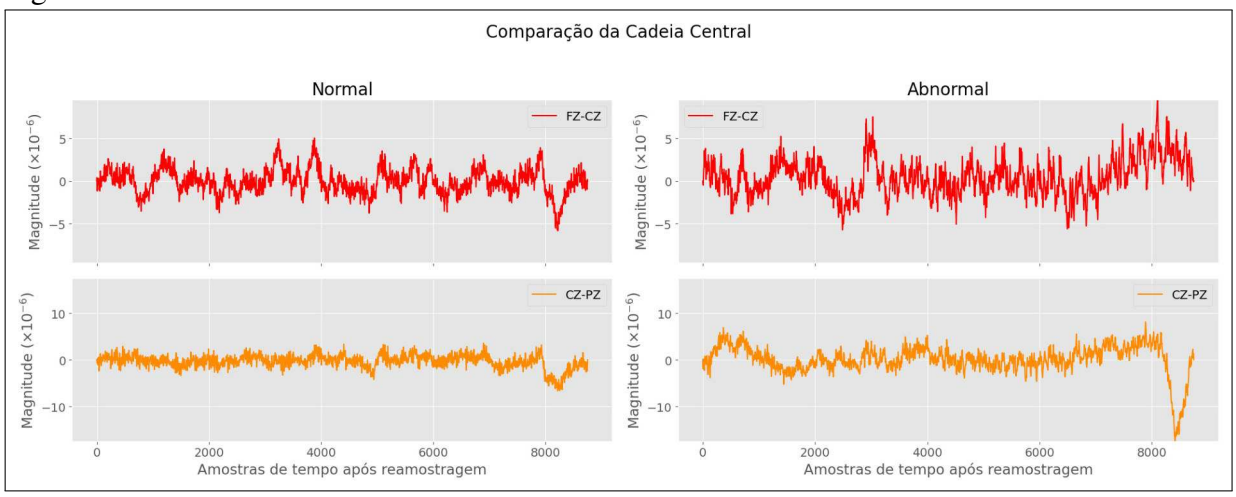
Fonte: elaborado pelo autor (2024).

Figura 23 – Cadeia Parassagital Direita.



Fonte: elaborado pelo autor (2024).

Figura 24 – Cadeia Central.



Fonte: elaborado pelo autor (2024).

4.2 Análise qualitativa - frequência e magnitude das ondas cerebrais

A obtenção das magnitudes e frequências das ondas cerebrais é feita usando Transformada de Fourier (outro método de análise de tempo-frequência, como a transformada *wavelet*), a qual decompõe os sinais da gravação em respectivas bandas de frequência explicitadas na seção da Fundamentação Teórica.

Com o que está documentado na literatura, os profissionais de saúde conseguem fundamentar seus diagnósticos baseando-se na comparação das magnitudes e frequências de cada banda com o padrão considerado saudável. Uma breve análise nesse sentido será adicionada antes da plotagem das figuras.

Relacionando com as doenças, em casos de epilepsia, a magnitude das bandas cerebrais aumenta durante as convulsões, especialmente nas bandas *theta* e *delta*, refletindo o impacto neuronal característico dessa condição. No caso de um acidente vascular cerebral (AVC), a magnitude tende a diminuir na região afetada, pois há perda de função neuronal e há redução da atividade elétrica causada pela interrupção do fluxo de sangue. E após uma concussão, observa-se uma redução na magnitude das ondas *beta*, indicando desaceleração da atividade cerebral, enquanto as ondas *theta* e *delta* podem aumentar, refletindo a desorganização neuronal em resposta ao trauma.

Sobre as frequências das bandas cerebrais, em casos de epilepsia, durante as convulsões, há grande presença das ondas lentas, predominantemente a *delta*, há também redução das ondas *alpha* e picos anormais. No contexto de AVC, há também aumento das ondas lentas, como *delta* e *theta*, assim como diminuição nas ondas *alpha* no local onde houve impedimento do fluxo de sangue. Já em casos de concussão, há aumento na proporção de ondas *delta* e *theta*, e diminuição intermitente das ondas *alpha*. (ATTAR, 2022)

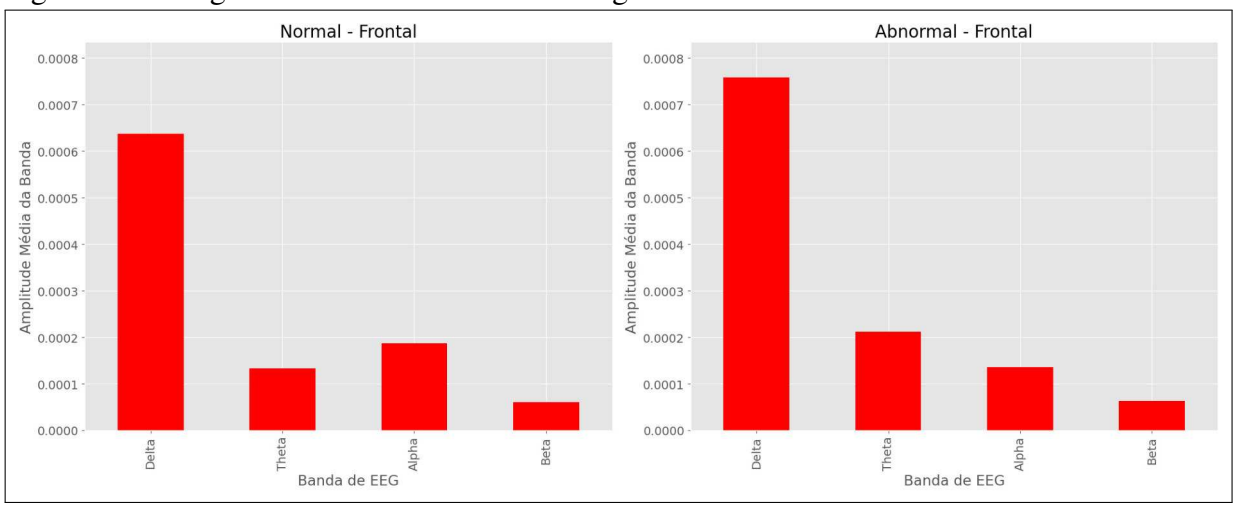
4.2.1 Análise e Visualização

Da mesma forma utilizada na plotagem '*double-banana*', foi feito o *plot* comparativo entre as classes para as ondas cerebrais *delta*, *theta*, *alpha* e *beta*, considerando o sinal médio de cada classe. Esse plot abrange tanto o cálculo da magnitude quanto da frequência que é possível se obter com a Transformada de Fourier do sinal.

Observando os gráficos comparativos das magnitudes por região, é possível observar as seguintes mudanças:

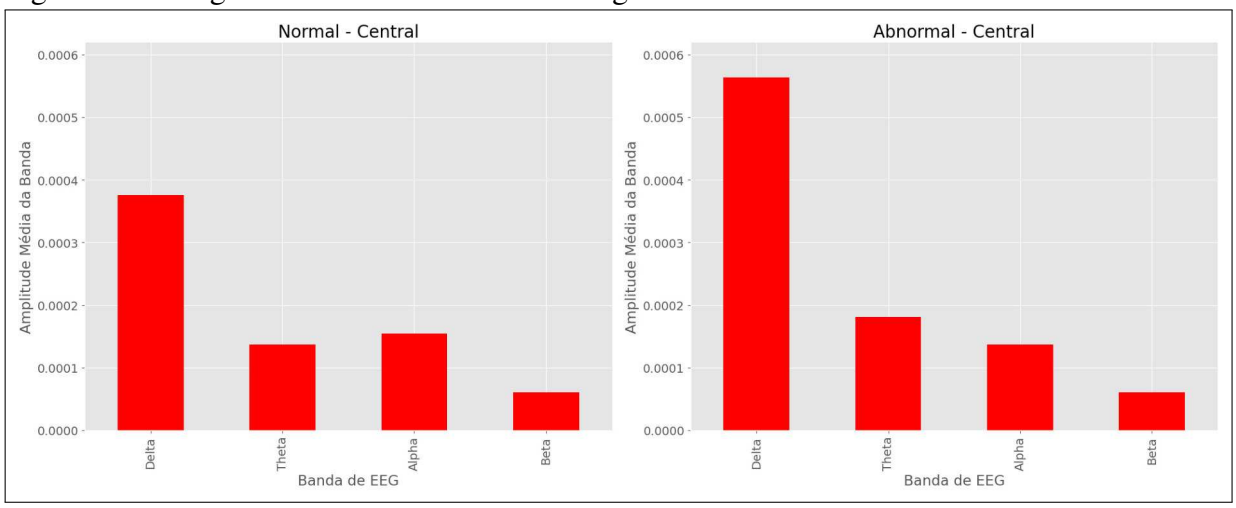
- Na região frontal, Figura 25, há aumento na banda *delta* o que pode indicar problemas originados de concussão, os quais podem promover anormalidades no controle emocional e comportamental do indivíduo;
- Na região central (compreendida entre as regiões frontal e parietal), Figura 26, há aumento na banda *delta*, que pode estar relacionado a problemas na coordenação motora;
- Na região temporal, Figura 27, há um aumento na banda *delta*, o que pode estar relacionado à problemas de memória e de audição;
- Na região parietal, Figura 28, o aumento considerável nas bandas *delta* e *theta* pode indicar problemas de causas diversas, mas danos nessa região podem resultar em problemas nos sentidos e orientação espacial;
- Na região ocipital, Figura 29, além do aumento na banda *delta*, há também diminuição leve na banda *alpha*. Isso pode estar relacionado com problemas visuais.

Figura 25 – Magnitude das ondas cerebrais - região frontal.



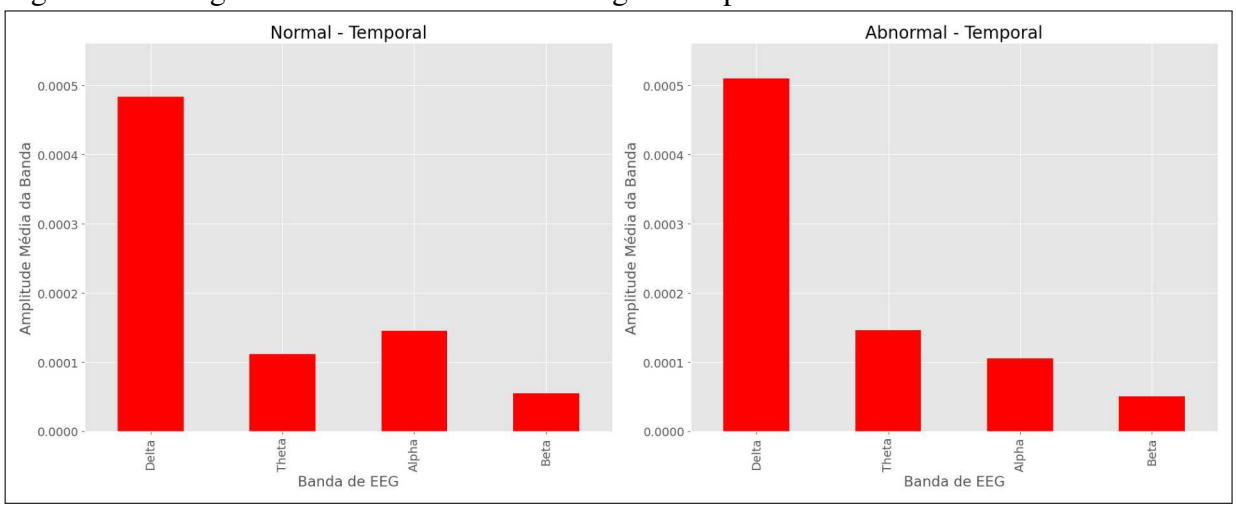
Fonte: elaborado pelo autor (2024).

Figura 26 – Magnitude das ondas cerebrais - região central.



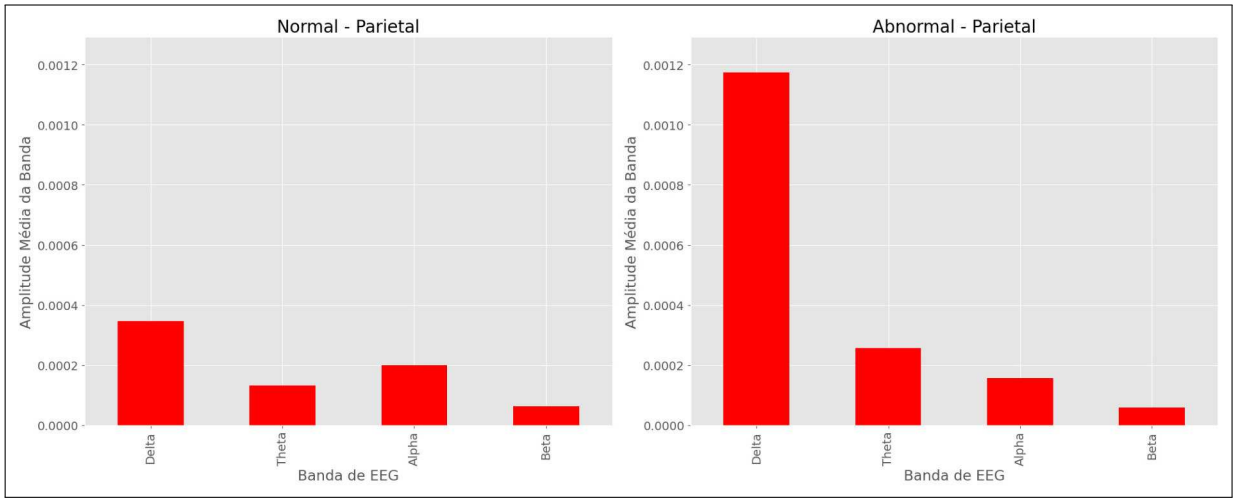
Fonte: elaborado pelo autor (2024).

Figura 27 – Magnitude das ondas cerebrais - região temporal.



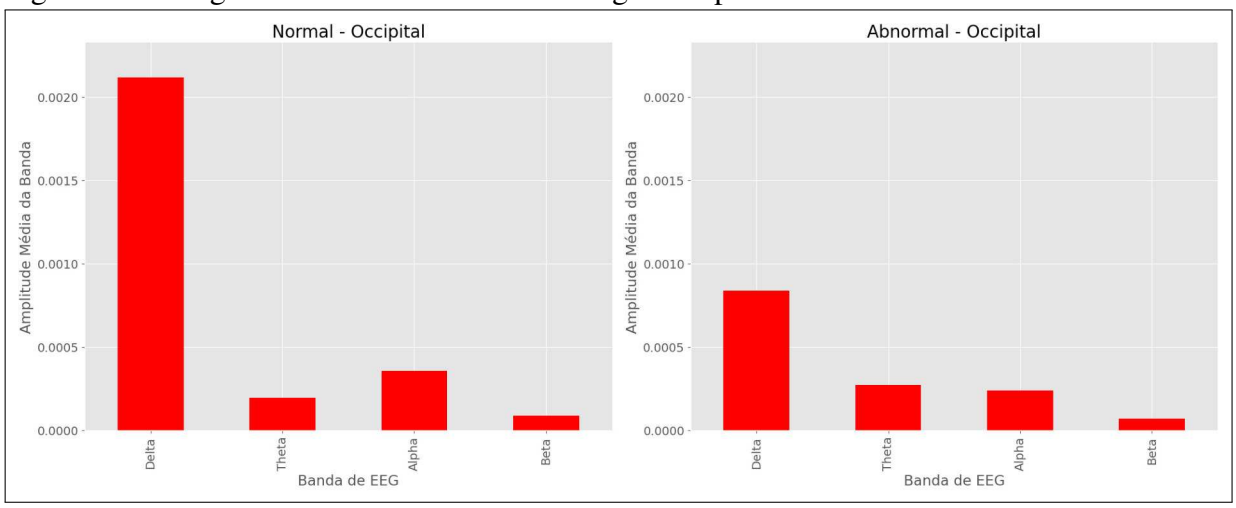
Fonte: elaborado pelo autor (2024).

Figura 28 – Magnitude das ondas cerebrais - região parietal.



Fonte: elaborado pelo autor (2024).

Figura 29 – Magnitude das ondas cerebrais - região ocipital.

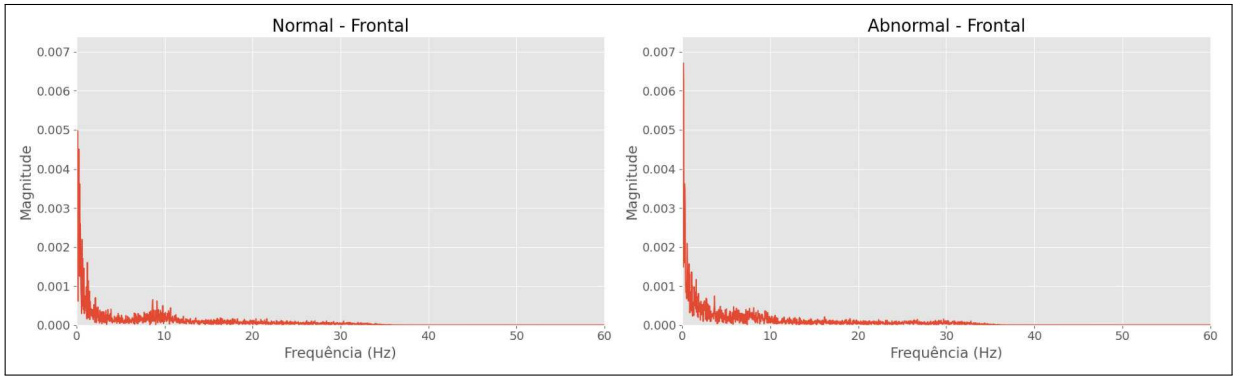


Fonte: elaborado pelo autor (2024).

Já sobre os gráficos comparativos das frequências por região, é possível observar as seguintes mudanças:

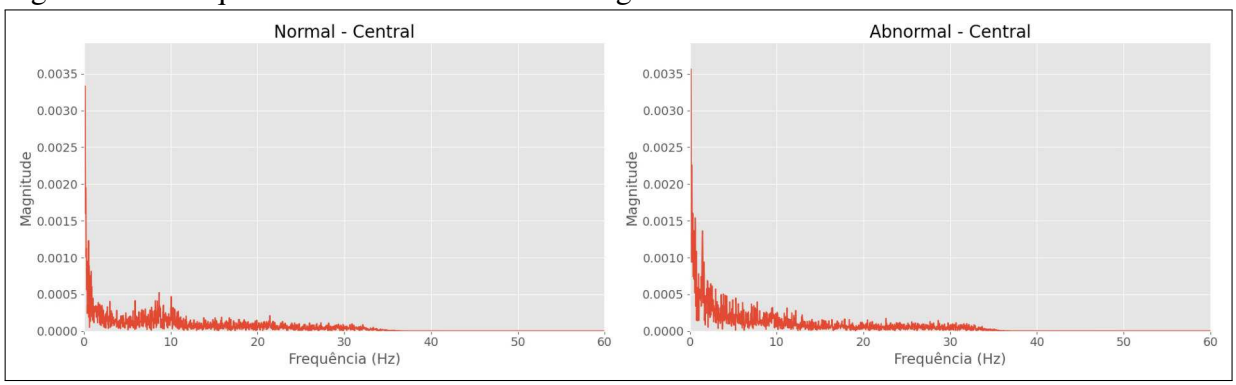
- Na região frontal, Figura 30, os gráficos também mostram um perfil semelhante com alta atividade na banda delta, tendo uma leve redução nas frequências mais altas na condição *abnormal*, indicando potenciais impactos nas funções cognitivas e comportamentais;
- Na região central, 31, há uma considerável redução na magnitude das bandas, o que sugere uma diminuição da atividade elétrica cerebral;
- Na região temporal, Figura 32, há uma redução diversa na magnitude de todas as bandas, o que pode indicar problemas na memória e na rede auditiva;
- Na região parietal, Figura 33, é visto que o nível de frequência para a classe *abnormal* está mais achatado, indicando as habilidades sensoriais e orientação espacial;
- Na região ocipital, Figura 34, há uma diminuição na atividade *delta*, o que pode estar relacionado a problemas na habilidade visual.

Figura 30 – Frequência das ondas cerebrais - região frontal.



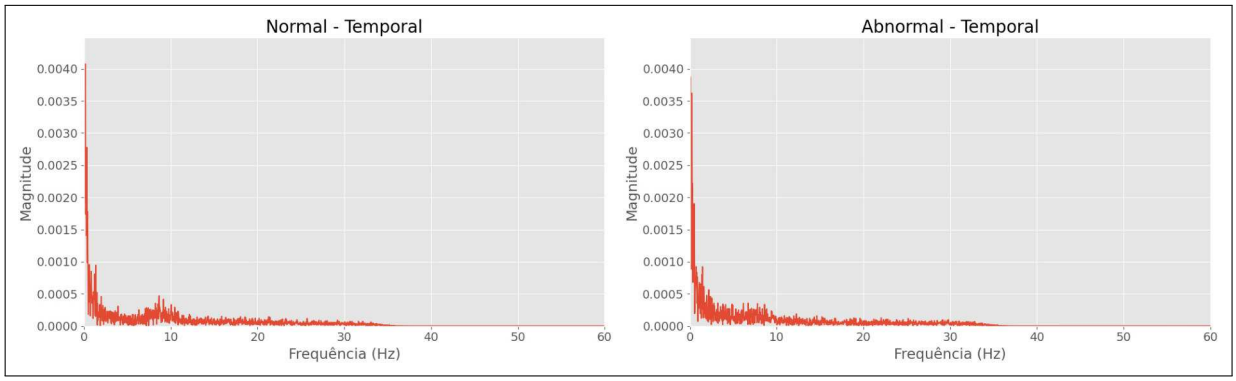
Fonte: elaborado pelo autor (2024).

Figura 31 – Frequência das ondas cerebrais - região central.



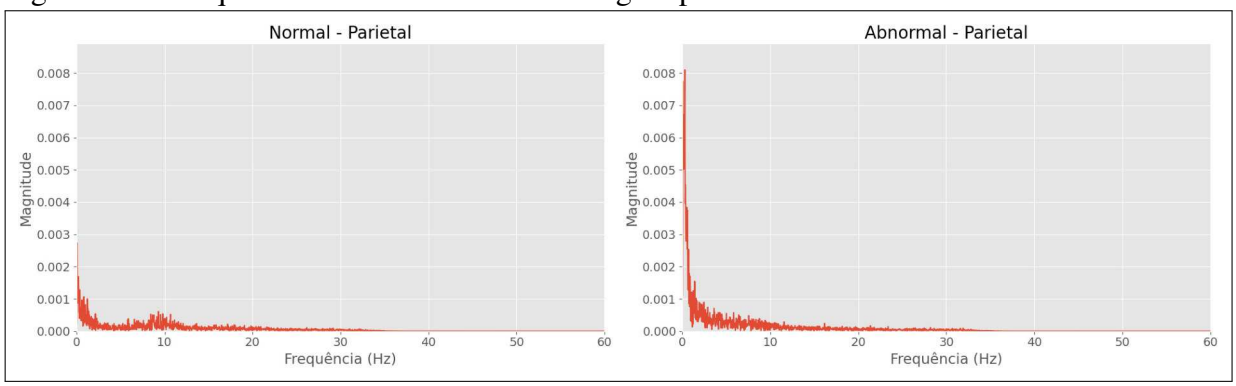
Fonte: elaborado pelo autor (2024).

Figura 32 – Frequência das ondas cerebrais - região temporal.



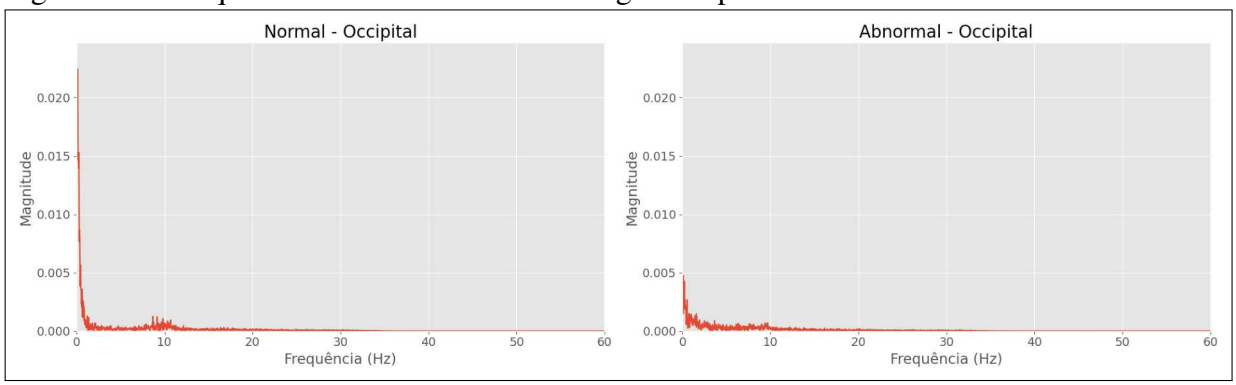
Fonte: elaborado pelo autor (2024).

Figura 33 – Frequência das ondas cerebrais - região parietal.



Fonte: elaborado pelo autor (2024).

Figura 34 – Frequência das ondas cerebrais - região ocipital.



Fonte: elaborado pelo autor (2024).

4.3 Análise quantitativa - etapa de classificação

Nesta seção será apresentada a análise quantitativa das métricas relacionadas à primeira parte do experimento principal: uso de modelos de classificação para gerar as métricas e comparar as performances. Os modelos, incluindo Random Forest, XGBoost, Gradient Boosting, K-Neighbors e Decision Tree, foram escolhidos devido aos algoritmos diferentes e, portanto, com complexidade variável e capacidade em lidar com dados de alta dimensionalidade. Relembrando, o procedimento envolveu variar o modelo de clusterização e o número de microestados para gerar os *datasets* de entrada, contendo os valores do GEV, ocorrência, duração e cobertura, os quais foram concatenados para gerar um novo *dataset* contendo dados relacionados a ambas as classes. Depois foi feita a separação entre o conjunto de treino e teste, variando este entre 10% a 90% do conjunto original, e, em seguida, foram aplicados os modelos de classificação, sendo feito o registro das métricas relacionadas, totalizando 8912 resultados para cada um dos cinco algoritmos de clusterização usados para geração do *dataset* de entrada.

É válido destacar o uso do parâmetro *random state* na função de separação do *dataset* concatenado para garantir a reprodutibilidade do experimento. Com esse parâmetro fixado, garante-se a consistência na execução da separação, eliminando variação nos resultados que poderiam ser causadas por diferentes amostragens dos dados.

4.3.1 Resultados considerando 20% do conjunto de dados para teste

Das planilhas de resultados geradas após a etapa de classificação, foi feita uma extração dos registros que obtiveram valores das métricas acurácia e precisão acima de 70% (Tabela 12). Essa escolha foi devido à performance geral do modelo e à necessidade de evitar

tratamentos desnecessários originados de diagnósticos incorretos: a acurácia permite uma avaliação geral do modelo e a precisão minimiza a ocorrência de falsos positivos, ou seja, considerando acertadamente a maioria dos pacientes identificados com doença.

O modelo *K-Neighbors* com 10 microestados e o algoritmo *kmedoids* destacou-se, alcançando a maior precisão (0,8333) e uma acurácia de 0,8. Já o modelo *OPF-chebyshev* com 16 microestados, utilizando o algoritmo *modified k-means*, mostrou um bom equilíbrio entre precisão (0,7) e *recall*, com uma acurácia também de 0,8. Isso sugere que os modelos são robustos e conseguem generalizar bem, mesmo em conjuntos de teste mais complexos.

Nota-se também que os algoritmos *K-Means* e *DBSCAN* não produziram resultados satisfatórios, sugerindo que não são adequados para a segmentação dos microestados de EEG neste contexto. Relembrando que *k-means*, que assume clusters esféricos de tamanho semelhante, pode ter falhado devido à complexidade dos dados. Já o *DBSCAN*, que se baseia na densidade, pode ter encontrado dificuldades em identificar clusters em um conjunto com alta variabilidade, que é típico de sinais de EEG. Por isso, algoritmos como *kmedoids* e *k-means* modificado tiveram um desempenho superior, pois são mais robustos frente a *outliers* e a diferentes formas dos clusters.

Tabela 12 – Resultados das métricas acurácia e precisão acima de 70%.

Modelo	**Núm. Microestados**	**Algoritmo**	**Acurácia**	**Precisão**
K-Neighbors	10	kmedoids	0,8	0,833333333
OPF-chebyshev	14	modified k-means	0,8	0,75
OPF-chebyshev	16	modified k-means	0,8	0,7
OPF-pearson	10	kmedoids	0,75	0,8
Gradient Boosting	6	aahc	0,75	0,8
OPF-additive_symmetric	7	kmedoids	0,75	0,714285714
OPF-kullback_leibler	10	kmedoids	0,75	0,714285714
Decision Tree	6	aahc	0,75	0,714285714
Gradient Boosting	2	aahc	0,75	0,714285714

Fonte: elaborado pelo autor (2024).

A Tabela 13 apresenta os tempos de teste para diferentes modelos. Observa-se que o modelo *K-Neighbors* com 10 microestados e utilizando o algoritmo *kmedoids* teve um dos tempos de teste mais rápidos, cerca de 0,014 segundos. Esse tempo relativamente baixo pode ser atribuído à simplicidade do algoritmo *kmedoids* em comparação com outros modelos mais complexos.

Por outro lado, modelos como *OPF-chebyshev* com 14 e 16 microestados, que utilizam o algoritmo *modified k-means*, apresentam tempos de teste consideravelmente maiores,

aproximadamente 0,028 e 0,019 segundos, respectivamente. Esse aumento no tempo de teste pode ser explicado pela complexidade adicional do *modified k-means*, que requer mais iterações e cálculos para convergir.

Modelos que utilizam o algoritmo *AAHC*, como *Gradient Boosting* e *Decision Tree*, exibem tempos de teste muito baixos, próximos de 0,001 a 0,002 segundos. Esses tempos rápidos indicam que, apesar da complexidade potencial dos modelos de *Gradient Boosting* e *Decision Tree*, o processo de teste é eficiente, possivelmente devido à forma como o *aahc* agrupa os dados de maneira hierárquica e otimizada.

Tabela 13 – Resultados de tempo de teste.

Modelo	**Núm. Microestados**	**Algoritmo**	**Tempo de Teste (s)**
K-Neighbors	10	kmedoids	0,014367342
OPF-chebyshev	14	modified k-means	0,028836489
OPF-chebyshev	16	modified k-means	0,019856215
OPF-pearson	10	kmedoids	0,027000189
Gradient Boosting	6	aahc	**0,007977724**
OPF-additive_symmetric	7	kmedoids	0,043694019
OPF-kullback_leibler	10	kmedoids	0,0306077
Decision Tree	6	aahc	**0,002316952**
Gradient Boosting	2	aahc	**0,001814127**

Fonte: elaborado pelo autor (2024).

As matrizes de confusão apresentadas (Tabelas 14, 15, 16, 17, 18, 19, 20, 21, 22), fornecem *insights* visuais acerca da performance dos modelos no contexto da análise EEG, considerando 20 gravações.

O modelo *K-Neighbors*, utilizando o algoritmo *kmedoids* com 10 microestados, mostrou um bom desempenho, especialmente na identificação de verdadeiros positivos, com 11 acertos. Isso sugere que o modelo tem uma alta capacidade de identificar corretamente as instâncias positivas, embora tenha tido alguns falsos positivos e falsos negativos.

Já o modelo *OPF-chebyshev* com 14 e 16 microestados, usando o algoritmo *modified k-means*, apresentou uma boa precisão, mas com uma ligeira queda no *recall*, indicando que o modelo teve dificuldade em identificar alguns verdadeiros negativos. No entanto, o balanceamento entre as métricas de precisão e *recall* demonstra uma robustez razoável do modelo.

O modelo *OPF-pearson* com *kmedoids* e 10 microestados mostrou-se eficaz na identificação de verdadeiros positivos, mas apresentou algumas dificuldades em identificar corretamente os verdadeiros negativos, como mostrado pelos 4 falsos negativos na Tabela.

O modelo *Gradient Boosting*, usando o algoritmo *aahc* com 6 e 2 microestados,

apresentou uma performance interessante. Especialmente com 6 microestados, o modelo conseguiu identificar uma boa quantidade de verdadeiros positivos, mas teve dificuldade com alguns verdadeiros negativos, sugerindo que o modelo pode estar um pouco mais inclinado a classificar instâncias como positivas.

Os modelos *OPF-additive_symmetric* e *OPF-kullback_leibler*, ambos utilizando o algoritmo *kmedoids*, tiveram desempenhos semelhantes, com bons valores de verdadeiros positivos, mas enfrentaram desafios em evitar falsos negativos. Isso pode indicar que esses modelos e algoritmos, embora eficientes, ainda têm espaço para melhorias na identificação correta dos casos negativos de doença.

Por fim, o modelo *Decision Tree* com 6 microestados usando *aahc* também mostrou um equilíbrio entre identificar verdadeiros positivos e verdadeiros negativos, mas assim como outros modelos, enfrentou dificuldades com alguns falsos negativos, sugerindo que talvez ajustes adicionais no algoritmo poderiam melhorar o desempenho global.

Tabela 14 – Matriz de Confusão para K-Neighbors com 10 microestados usando kmedoids.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	11	1
Verdadeiro Negativo	3	5

Fonte: elaborado pelo autor (2024).

Tabela 15 – Matriz de Confusão para OPF-*chebyshev* com 14 microestados usando modified k-means.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	10	2
Verdadeiro Negativo	2	6

Fonte: elaborado pelo autor (2024).

Tabela 16 – Matriz de Confusão para OPF-*chebyshev* com 16 microestados usando modified k-means.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	9	3
Verdadeiro Negativo	1	7

Fonte: elaborado pelo autor (2024).

Tabela 17 – Matriz de Confusão para OPF-*pearson* com 10 microestados usando kmedoids.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	11	1
Verdadeiro Negativo	4	4

Fonte: elaborado pelo autor (2024).

Tabela 18 – Matriz de Confusão para Gradient Boosting com 6 microestados usando AAHC.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	11	1
Verdadeiro Negativo	4	4

Fonte: elaborado pelo autor (2024).

Tabela 19 – Matriz de Confusão para OPF-*additive_symmetric* com 7 microestados usando kmedoids.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	10	2
Verdadeiro Negativo	3	5

Fonte: elaborado pelo autor (2024).

Tabela 20 – Matriz de Confusão para OPF-*kullback_leibler* com 10 microestados usando kmedoids.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	10	2
Verdadeiro Negativo	3	5

Fonte: elaborado pelo autor (2024).

Tabela 21 – Matriz de Confusão para Decision Tree com 6 microestados usando AAHC.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	10	2
Verdadeiro Negativo	3	5

Fonte: elaborado pelo autor (2024).

Tabela 22 – Matriz de Confusão para Gradient Boosting com 2 microestados usando AAHC.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	10	2
Verdadeiro Negativo	3	5

Fonte: elaborado pelo autor (2024).

4.3.2 Resultados considerando 30% do conjunto de dados para teste

Semelhante ao que foi feito na seção anterior, considerando 30% do conjunto completo para teste, será feita a análise comparativa dos valores de acurácia e precisão, do tempo de teste e das respectivas matrizes de confusão.

A Tabela 23 apresenta os resultados de acurácia e precisão para diferentes modelos de clusterização de microestados de EEG, com valores de ambas as métricas acima de 70%. Nota-se

que o modelo *OPF-bray_curtis*, utilizando o algoritmo *modified k-means*, obteve consistência nos resultados, com acurácia de 76,67% e precisão de 71,43%, assim como os demais modelos que também utilizam esse algoritmo, sugerindo uma robustez em seu desempenho.

Outro ponto que merece destaque é o desempenho do modelo *OPF-min_symmetric* com o algoritmo *kmeans*, que alcançou uma das mais altas precisões (85,71%) juntamente com uma acurácia de 73,33%, indicando uma boa capacidade desse modelo de evitar falsos positivos.

O algoritmo *DBSCAN* mostra novamente desempenho insuficiente, mesmo alcançando estabilidade com uma precisão constante de 70%. Isso retrata uma menor capacidade de diferenciar corretamente os casos positivos, quando comparado aos outros algoritmos.

Em resumo, a Tabela sugere que algoritmos como *modified k-means* e *k-means* são mais adequados para a segmentação de microestados de EEG. Por outro lado, o *DBSCAN* mostrou-se novamente menos eficaz nesse contexto específico, apresentando resultados consistentes, mas com precisão um pouco inferior.

Tabela 23 – Resultados das métricas de acurácia e precisão acima de 70%.

Modelo	**Núm. Microestados**	**Algoritmo**	**Acurácia**	**Precisão**
OPF-bray_curtis	16	modified k-means	0,7667	0,7143
OPF-gower	16	modified k-means	0,7667	0,7143
OPF-kulczynski	16	modified k-means	0,7667	0,7143
OPF-manhattan	16	modified k-means	0,7667	0,7143
OPF-non_intersection	16	modified k-means	0,7667	0,7143
OPF-soergel	16	modified k-means	0,7667	0,7143
OPF-hamming	5	kmedoids	0,7333	1
OPF-min_symmetric	18	kmeans	0,7333	0,8571
OPF-statistic	16	kmedoids	0,7333	0,7778
OPF-statistic	14	kmeans	0,7	1
OPF-statistic	19	aahc	0,7	1
OPF-statistic	12	kmeans	0,7	0,8333
OPF-statistic	16	kmeans	0,7	0,8333
OPF-mean_censored_euclidean	14	dbscan	0,7	0,8333
OPF-min_symmetric	17	kmeans	0,7	0,75
OPF-mean_censored_euclidean	8	dbscan	0,7	0,7
OPF-mean_censored_euclidean	9	dbscan	0,7	0,7
OPF-hamming	18	aahc	0,7	0,7

Fonte: elaborado pelo autor (2024).

A Tabela 24 mostra os tempos de teste para diferentes modelos e algoritmos utilizados na segmentação de microestados de EEG. Observa-se que o modelo OPF-gower com 16 microestados, utilizando o algoritmo *modified k-means*, teve o menor tempo de teste (0,02376914 s), seguido de perto pelos modelos OPF-kulczynski e OPF-manhattan, que também utilizaram o *modified k-means* e apresentaram tempos de teste semelhantes. Esses resultados indicam que o

algoritmo *modified k-means*, quando aplicado a esses modelos, não só é eficiente em termos de precisão, como também em termos de tempo computacional.

Por outro lado, o modelo OPF-mean_censored_euclidean com 9 microestados, utilizando o algoritmo *DBSCAN*, apresentou um dos maiores tempos de teste (0,062419891 s), sugerindo que o *DBSCAN* pode demandar mais tempo para processar os dados com baixa densidade.

Tabela 24 – Resultados de tempo de teste para diferentes modelos e algoritmos.

Modelo	Núm. Microestados	Algoritmo	Tempo de Teste (s)
OPF-bray_curtis	16	*modified k-means*	0,039472818
OPF-gower	16	*modified k-means*	**0,02376914**
OPF-kulczynski	16	*modified k-means*	0,037503481
OPF-manhattan	16	*modified k-means*	**0,025009155**
OPF-non_intersection	16	*modified k-means*	0,025421143
OPF-soergel	16	*modified k-means*	0,040947676
OPF-hamming	5	*kmedoids*	0,020559549
OPF-min_symmetric	18	*kmeans*	0,044415712
OPF-statistic	16	*kmedoids*	**0,017991304**
OPF-statistic	14	*kmeans*	0,036516666
OPF-statistic	19	*aahc*	0,049479723
OPF-statistic	12	*kmeans*	0,029304981
OPF-statistic	16	*kmeans*	0,043455124
OPF-mean_censored_euclidean	14	*dbscan*	0,049389124
OPF-min_symmetric	17	*kmeans*	0,061274767
OPF-mean_censored_euclidean	8	*dbscan*	0,062419891
OPF-mean_censored_euclidean	9	*dbscan*	0,050566912
OPF-hamming	18	*aahc*	0,030025721

Fonte: elaborado pelo autor (2024).

As matrizes de confusão apresentadas nas Tabelas (25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42), considerando 20 gravações, mostram que, em geral, os modelos de microestados de EEG utilizando o algoritmo *modified k-means* demonstraram um bom equilíbrio entre verdadeiros positivos e negativos, mas ainda enfrentam alguns desafios em evitar falsos negativos, especialmente nos casos com maior número de microestados. Modelos como o OPF-*bray_curtis* e OPF-*gower* mantiveram uma performance estável, enquanto o OPF-*hamming* com 5 microestados utilizando o algoritmo *kmedoids* apresentou uma alta taxa de verdadeiros positivos, mas teve dificuldade em minimizar os falsos negativos. Por outro lado, os modelos usando *k-means* mostraram variações na precisão dependendo do número de microestados, sugerindo que ajustes adicionais podem ser necessários para otimizar a identificação dos microestados corretos.

Tabela 25 – Matriz de Confusão para OPF-*bray_curtis* com 16 microestados usando *modified k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	13	4
Verdadeiro Negativo	3	10

Fonte: elaborado pelo autor (2024).

Tabela 26 – Matriz de Confusão para OPF-*gower* com 16 microestados usando *modified k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	13	4
Verdadeiro Negativo	3	10

Fonte: elaborado pelo autor (2024).

Tabela 27 – Matriz de Confusão para OPF-*kulczynski* com 16 microestados usando *modified k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	13	4
Verdadeiro Negativo	3	10

Fonte: elaborado pelo autor (2024).

Tabela 28 – Matriz de Confusão para OPF-*manhattan* com 16 microestados usando *modified k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	13	4
Verdadeiro Negativo	3	10

Fonte: elaborado pelo autor (2024).

Tabela 29 – Matriz de Confusão para OPF-*non_intersection* com 16 microestados usando *modified k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	13	4
Verdadeiro Negativo	3	10

Fonte: elaborado pelo autor (2024).

Tabela 30 – Matriz de Confusão para OPF-*soergel* com 16 microestados usando *modified k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	13	4
Verdadeiro Negativo	3	10

Fonte: elaborado pelo autor (2024).

Tabela 31 – Matriz de Confusão para OPF-*hamming* com 5 microestados usando *kmedoids*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	17	0
Verdadeiro Negativo	8	5

Fonte: elaborado pelo autor (2024).

Tabela 32 – Matriz de Confusão para OPF-*min_symmetric* com 18 microestados usando *k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	16	1
Verdadeiro Negativo	7	6

Fonte: elaborado pelo autor (2024).

Tabela 33 – Matriz de Confusão para OPF-*statistic* com 16 microestados usando *kmedoids*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	15	2
Verdadeiro Negativo	6	7

Fonte: elaborado pelo autor (2024).

Tabela 34 – Matriz de Confusão para OPF-*statistic* com 14 microestados usando *k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	17	0
Verdadeiro Negativo	9	4

Fonte: elaborado pelo autor (2024).

Tabela 35 – Matriz de Confusão para OPF-*statistic* com 19 microestados usando *AAHC*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	17	0
Verdadeiro Negativo	9	4

Fonte: elaborado pelo autor (2024).

Tabela 36 – Matriz de Confusão para OPF-*statistic* com 12 microestados usando *k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	16	1
Verdadeiro Negativo	8	5

Fonte: elaborado pelo autor (2024).

Tabela 37 – Matriz de Confusão para OPF-*statistic* com 16 microestados usando *k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	16	1
Verdadeiro Negativo	8	5

Fonte: elaborado pelo autor (2024).

Tabela 38 – Matriz de Confusão para OPF-*mean_censored_euclidean* com 14 microestados usando *DBSCAN*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	16	1
Verdadeiro Negativo	8	5

Fonte: elaborado pelo autor (2024).

Tabela 39 – Matriz de Confusão para OPF-*min_symmetric* com 17 microestados usando *k-means*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	15	2
Verdadeiro Negativo	7	6

Fonte: elaborado pelo autor (2024).

Tabela 40 – Matriz de Confusão para OPF-*mean_censored_euclidean* com 8 microestados usando *DBSCAN*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	14	3
Verdadeiro Negativo	6	7

Fonte: elaborado pelo autor (2024).

Tabela 41 – Matriz de Confusão para OPF-*mean_censored_euclidean* com 9 microestados usando *DBSCAN*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	14	3
Verdadeiro Negativo	6	7

Fonte: elaborado pelo autor (2024).

Tabela 42 – Matriz de Confusão para OPF-*hamming* com 18 microestados usando *AAHC*.

	Predito Positivo	Predito Negativo
Verdadeiro Positivo	14	3
Verdadeiro Negativo	6	7

Fonte: elaborado pelo autor (2024).

4.3.3 Melhores resultados no geral

A Tabela 43 apresenta os melhores resultados obtidos utilizando o algoritmo *k-means* para diferentes modelos de clusterização de microestados. Todos os modelos listados demonstraram alta acurácia (0.9) e, na maioria dos casos, precisões perfeitas (1.0), o que indica que esses modelos foram eficazes em identificar corretamente os casos positivos. No entanto, o *Recall* variou entre 0.75 e 1, sugerindo que alguns modelos, embora precisos, não conseguiram identificar todos os casos positivos.

O modelo OPF-*pearson*, por exemplo, apresentou um Revocação de 0.75, o que significa que ele falhou em capturar 25% dos casos positivos, apesar de sua alta precisão. Por outro lado, modelos como o OPF-*min_symmetric* e OPF-*chebyshev* mostraram uma combinação mais equilibrada entre precisão e *Recall*, refletida em um *F1 Score* de 0.88.

Tabela 43 – Melhores resultados utilizando o algoritmo *kmeans*.

Modelo	**Microestados**	**Tempo de Teste (s)**	**Acurácia**	**Precisão**	***Revocação***	***F1 Score***
OPF-pearson	10	0,023843765	0,9	1	0,75	0,85
OPF-clark	19	0,035444498	0,9	1	0,75	0,85
OPF-min_symmetric	11	0,030693531	0,9	0,8	1	0,88
OPF-min_symmetric	14	0,028050184	0,9	0,8	1	0,88
OPF-min_symmetric	15	0,026992559	0,9	0,8	1	0,88
OPF-chebyshev	18	0,015526533	0,9	0,8	1	0,88

Fonte: elaborado pelo autor (2024).

Na Tabela 44, o desempenho perfeito alcançado pelo modelo *K-Neighbors* é um resultado que sugere *a priori* que o modelo foi altamente eficaz em capturar os padrões corretos durante o treinamento. No entanto, é importante considerar a possibilidade de que o modelo esteja se beneficiando de certas características dos dados que podem não estar presentes em outras situações, por exemplo, a quantidade e o tamanho de cada amostra nos conjuntos de treino e teste. Como forma de atestar a confiabilidade dos resultados, é possível aplicar técnicas de validação cruzada e, eventualmente, confirmar a robustez do modelo e capacidade de generalização para o problema.

Tabela 44 – Melhores resultados utilizando o algoritmo *kmedoids*.

Modelo	**Microestados**	**Tempo de Teste (s)**	**Acurácia**	**Precisão**	**Revocação**	***F1 Score***
K-Neighbors	10	0,005492449	1	1	1	1
OPF-canberra	13	0,022422075	0,9	0,8	1	0,888888889

Fonte: elaborado pelo autor (2024).

A Tabela 47 destaca uma variedade de modelos e microestados. Observa-se que o

modelo *Gradient Boosting* com 5 microestados atingiu um desempenho perfeito, com acurácia, precisão, revocação e *F1 Score* todas iguais a 1, evidenciando uma excelente capacidade de classificação, com ressalva à questão do *overfitting* comentada no parágrafo referente a Tabela 44. Outros modelos, como o *XGBoost* e o *OPF-min_symmetric*, também apresentaram resultados robustos, com acurácia de 0,9 e *F1 Scores* acima de 0,85. Esses resultados indicam que o *modified k-means* é um algoritmo eficaz para a clusterização de microestados, mas o desempenho pode variar significativamente dependendo do modelo e do número de microestados selecionados.

Além disso, o modelo *Gradient Boosting* com 5 microestados destacou-se por apresentar o menor tempo de teste, de apenas 0,002701283 segundos, demonstrando não apenas alta precisão, mas também uma eficiência excepcional. Em contraste, o *XGBoost* com 15 microestados registrou um dos maiores tempos de teste, atingindo 0,065442324 segundos, o que ressalta a importância de considerar o tempo de processamento na análise eficiente de exames EEG, especialmente em ambientes clínicos onde a rapidez é importante.

Tabela 45 – Melhores resultados utilizando o algoritmo *modified k-means*.

Modelo	**Microestados**	**Tempo de Teste (s)**	**Acurácia**	**Precisão**	**Revocação**	***F1 Score***
Gradient Boosting	5	0,002701283	1	1	1	1
XGBoost	6	0,008732319	0,9	1	0,75	0,85
OPF-min_symmetric	12	0,018820286	0,9	1	0,75	0,85
XGBoost	14	0,027873278	0,9	1	0,75	0,85
OPF-lorentzian	16	0,02380228	0,9	1	0,75	0,85
Random Forest	5	0,006985903	0,9	0,8	1	0,88
OPF-canberra	5	0,029198647	0,9	0,8	1	0,88
OPF-pearson	12	0,021104336	0,9	0,8	1	0,88
Gradient Boosting	14	0,005399227	0,9	0,8	1	0,88
Random Forest	18	0,01568079	0,9	0,8	1	0,88
Naive Bayes	19	0,00489068	0,9	0,8	1	0,88
Random Forest	3	0,007508039	0,8	0,75	0,75	0,75
Decision Tree	4	0,00170064	0,8	0,75	0,75	0,75
OPF-average_euclidean	4	0,019424677	0,8	0,75	0,75	0,75
OPF-chord	4	0,020091772	0,8	0,75	0,75	0,75
OPF-cosine	4	0,026745796	0,8	0,75	0,75	0,75
OPF-dice	4	0,0178473	0,8	0,75	0,75	0,75
OPF-euclidean	4	0,020135641	0,8	0,75	0,75	0,75
OPF-jaccard	4	0,056808472	0,8	0,75	0,75	0,75
OPF-log_euclidean	4	0,017169714	0,8	0,75	0,75	0,75
OPF-log_squared_euclidean	4	0,016603231	0,8	0,75	0,75	0,75
OPF-squared_euclidean	4	0,018040419	0,8	0,75	0,75	0,75
OPF-vicis_wave_hedges	5	0,023303986	0,8	0,75	0,75	0,75
OPF-bray_curtis	8	0,02393055	0,8	0,75	0,75	0,75
OPF-gower	8	0,014861107	0,8	0,75	0,75	0,75
OPF-hassanat	8	0,022671461	0,8	0,75	0,75	0,75
OPF-kulczynski	8	0,017745256	0,8	0,75	0,75	0,75
OPF-manhattan	8	0,015153646	0,8	0,75	0,75	0,75
OPF-non_intersection	8	0,016878366	0,8	0,75	0,75	0,75
OPF-soergel	8	0,024832249	0,8	0,75	0,75	0,75
Random Forest	13	0,012490511	0,8	0,75	0,75	0,75
OPF-chebyshev	14	0,01611495	0,8	0,75	0,75	0,75

Fonte: elaborado pelo autor (2024).

A Tabela 48 também apresenta uma variedade de modelos com diferentes números de microestados. O modelo *XGBoost* com 2 microestados obteve um *F1 Score* de 0,85, com alta precisão e revocação. Outros modelos, como *Decision Tree* e *Gradient Boosting*, também mostraram desempenhos significativos, com *F1 Scores* de 0,88. A acurácia variou entre 0,8 e 0,9, indicando uma boa performance geral do algoritmo. O tempo de teste variou consideravelmente entre os modelos, com destaque para o modelo *XGBoost* com 2 microestados, que apresentou o menor tempo de teste, de apenas 0,008259058 segundos. Em contraste, o modelo *XGBoost* com 15 microestados teve um dos maiores tempos de teste, atingindo 0,065442324 segundos.

Tabela 46 – Melhores resultados utilizando o algoritmo AAHC.

Modelo	**Microestados**	**Tempo de Teste (s)**	**Acurácia**	**Precisão**	**Revocação**	***F1 Score***
XGBoost	2	0,008259058	0,9	1	0,75	0,85
Decision Tree	18	0,006130934	0,9	0,8	1	0,88
Gradient Boosting	18	0,004074097	0,9	0,8	1	0,88
Gradient Boosting	15	0,003214121	0,8	0,75	0,75	0,75
XGBoost	15	0,065442324	0,8	0,75	0,75	0,75
OPF-bray_curtis	15	0,028116465	0,8	0,75	0,75	0,75
OPF-chi_squared	15	0,018884659	0,8	0,75	0,75	0,75
OPF-gower	15	0,014650583	0,8	0,75	0,75	0,75
OPF-hellinger	15	0,016130447	0,8	0,75	0,75	0,75
OPF-jeffreys	15	0,023011923	0,8	0,75	0,75	0,75
OPF-jensen	15	0,03295207	0,8	0,75	0,75	0,75
OPF-jensen_shannon	15	0,028085232	0,8	0,75	0,75	0,75
OPF-kulczynski	15	0,019806385	0,8	0,75	0,75	0,75
OPF-kullback_leibler	15	0,021493196	0,8	0,75	0,75	0,75
OPF-manhattan	15	0,014594793	0,8	0,75	0,75	0,75
OPF-matusita	15	0,01632309	0,8	0,75	0,75	0,75
OPF-non_intersection	15	0,016933918	0,8	0,75	0,75	0,75
OPF-sangvi	15	0,022801161	0,8	0,75	0,75	0,75
OPF-soergel	15	0,027973175	0,8	0,75	0,75	0,75
OPF-squared	15	0,026019573	0,8	0,75	0,75	0,75
OPF-squared_chord	15	0,019292831	0,8	0,75	0,75	0,75
OPF-topsoe	15	0,035200119	0,8	0,75	0,75	0,75
OPF-vicis_symmetric3	15	0,022063971	0,8	0,75	0,75	0,75
OPF-neyman	5	0,018269777	0,8	0,75	0,75	0,75
Random Forest	2	0,010499716	0,8	0,75	0,75	0,75
Decision Tree	2	0,001226187	0,8	0,75	0,75	0,75
Gradient Boosting	2	0,001820803	0,8	0,75	0,75	0,75
Naive Bayes	2	0,001458168	0,8	0,75	0,75	0,75
OPF-divergence	12	0.019395351	0,8	0,75	0,75	0,75

Fonte: elaborado pelo autor (2024).

Em suma, os métodos que mais se destacaram foram o *modified k-means* devido ao processo de clusterização considerar os centroides identificadores dos *clusters* como possivelmente virtuais, trazendo resistência aos *outliers* e à complexidade dos dados iniciais (robustez). Outro bom algoritmo foi o AAHC, trazendo as mesmas vantagens, mas com o processo de clusterização hierárquico, o que possibilita separar de forma mais granular ou específica os dados iniciais.

4.4 Análise quantitativa - etapa de clusterização

Neste tópico, foram analisados os algoritmos AAHC e *Modified K-Means*, que se destacaram com os melhores resultados na etapa de classificação. Utilizando métricas como *Silhouette*, Calinski-Harabasz, Davies-Bouldin e *Dunn Index*, avaliamos a qualidade da clusterização, com foco na separação inter e intra-cluster. O objetivo foi entender melhor como a qualidade da clusterização pode estar relacionada ao desempenho observado na classificação.

A Tabela 47 apresenta os resultados das métricas de clusterização para o algoritmo *Modified K-Means* com diferentes números de microestados, conforme discutido anteriormente. A métrica Silhouette, que avalia a separação entre clusters, obteve seus melhores valores com 3 e 4 microestados, embora os valores negativos indiquem uma separação subótima, possivelmente sugerindo sobreposição entre os *clusters*. Em contrapartida, a métrica Calinski-Harabasz, que favorece valores elevados, destacou 5 e 6 microestados, sugerindo uma maior densidade intra-*cluster* e uma separação inter-*cluster* mais eficiente. Já a métrica Davies-Bouldin, que prefere valores mais baixos, apontou 16 e 14 microestados como as melhores configurações, indicando uma separação mais clara entre os *clusters*. Por fim, a métrica Dunn, que também valoriza altos valores, confirmou 3 e 5 microestados como as configurações que proporcionaram uma separação relativa mais robusta entre os *clusters*.

Tabela 47 – Resultados das métricas de clusterização para o algoritmo *Modified K-means*.

Núm. Microestados	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**	**Dunn**
5	-0,0410576	**151,556956**	11,2844001	**0,00131516**
6	-0,0504862	**150,138618**	11,0267358	0,00115075
12	-0,0746555	136,412857	10,0991853	0,00035343
14	-0,0789003	138,468002	9,9601888	0,00024687
16	-0,0823895	138,321872	**9,8692306**	0,00026210
18	-0,0843572	133,167174	9,9742923	0,00034982
19	-0,0868675	133,396560	**9,8390562**	0,00035490
3	**-0,0090678**	142,323057	10,9161859	**0,00276025**
4	**-0,0296757**	149,952004	11,4265062	0,00119589
8	-0,0607445	148,068293	10,8133874	0,00098670
13	-0,0764664	138,026588	10,1660473	0,00043303

Fonte: Elaborado pelo autor (2024).

A Tabela 48 apresenta os resultados das métricas de clusterização para o algoritmo AAHC com diferentes números de microestados. Observa-se que a configuração com 2 microestados destacou-se em todas as métricas, com os melhores valores de *Silhouette*, Calinski-Harabasz, Davies-Bouldin e Dunn, indicando uma separação inter-cluster sólida e uma boa coesão intra-*cluster*. A configuração com 5 microestados também mostrou bons resultados, particularmente em *Silhouette* e Calinski-Harabasz, sugerindo que, além dos 2 microestados, esta pode ser uma configuração eficiente. Em contraste, as demais configurações apresentaram desempenho inferior, o que pode indicar que o aumento no número de microestados não necessariamente melhora a qualidade da clusterização. Esses resultados sugerem que, para o algoritmo AAHC, configurações com um menor número de microestados podem ser mais eficazes em termos de separação e coesão dos *clusters*.

Tabela 48 – Resultados das métricas de clusterização para o algoritmo AAHC.

Núm. Microestados	Silhouette	Calinski-Harabasz	Davies-Bouldin	Dunn
2	**0,905906073**	**159,8832767**	**0,055614962**	**1,787773912**
18	0,110677589	90,95343285	0,898960519	0,196039856
15	0,180535631	95,6416414	0,682860433	0,208673108
5	**0,561065645**	**114,4918928**	**0,357069696**	**0,393635072**
12	0,171720733	97,97426655	0,756600094	0,140056184

Fonte: Elaborado pelo autor (2024).

Em suma, a análise das tabelas comparando os algoritmos *Modified K-Means* e AAHC revela que o AAHC obteve um desempenho superior em termos de métricas de clusterização, especialmente quando utilizado com um menor número de microestados. O AAHC, com 2 microestados, destacou-se em todas as métricas, sugerindo uma melhor separação e coesão dos *clusters*, o que é crucial para uma análise eficiente dos padrões de EEG. Por outro lado, o *Modified K-Means* mostrou resultados mais dispersos, com diferentes números de microestados favorecendo métricas específicas, mas sem a consistência observada no AAHC. Isso indica que, enquanto o AAHC parece mais robusto e eficaz na configuração de 2 microestados, o *Modified K-Means* pode necessitar de ajustes mais cuidadosos no número de microestados para otimizar sua performance em diferentes contextos de análise.

Entretanto, é importante ressaltar que o uso de um número muito baixo de microestados, como 2, tende a não capturar a complexidade necessária dos sinais de EEG, conforme destacado nos trabalhos científicos. Vários estudos publicados sugerem que um número maior de microestados pode ser necessário para refletir de forma mais precisa a diversidade das dinâmicas cerebrais. Portanto, embora esses resultados com 2 microestados tenham apresentado bons valores para as métricas, é possível que essa configuração possa não ser suficiente para estudos mais detalhados, sendo necessária uma gama maior de experimentos para melhorar a qualidade da clusterização e corresponder à complexidade dos dados EEG frente ao número de microestados.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, como parte principal, foi feita uma análise ampla do comportamento de diferentes algoritmos de clusterização e classificação no contexto da análise de dados de EEG, com foco na comparação do número de *clusters* (microestados) e as respectivas métricas resultantes. Ao longo do estudo, foi verificado que, embora o número de *clusters* tradicionalmente utilizado em trabalhos acadêmicos apresente resultados robustos, o uso de outros números de *clusters* revelou potencial para melhorias nas métricas de desempenho dos algoritmos. As métricas de avaliação utilizadas corroboraram a identificação de um número ótimo de *clusters*, refletindo um desempenho aprimorado nos algoritmos de classificação, particularmente na distinção entre EEGs normais e anormais. Por fim, este estudo disponibilizou um indicativo técnico validado para a aplicação de diversas ferramentas de aprendizagem de máquina no contexto de EEG, contribuindo para avanços na utilização dessas tecnologias em diagnósticos clínicos.

Os experimentos também trouxeram análises clínicas comuns que se utilizam de inspeção visual para apoiar um diagnóstico, como a montagem bipolar longitudinal e a análise de amplitude e frequência das ondas cerebrais, porém, o foco do trabalho reside na avaliação das métricas frente aos diferentes algoritmos de clusterização e classificação, utilizando diversos números de *clusters*. Em destaque, os algoritmos de clusterização AAHC e *modified K-Means* apresentaram os melhores resultados de métricas na etapa de classificação, quando usados em conjunto dos melhores classificadores do estado da arte mais recente, *Gradient Boosting* e *XGBoost*, em termos de rapidez e consistência de convergência. Esse resultado também se apoia na análise da clusterização desses algoritmos usando métricas comuns de quantização da coesão dos elementos de um *cluster* e separação dos *clusters* entre si.

Outra informação obtida ao analisar os valores das métricas é que números maiores de microestados não necessariamente implicam em uma melhor configuração no uso de ferramentas de aprendizagem de máquina para estudar os microestados em EEG. Diante disso, vê-se que uma quantidade de microestados próxima ao comum da literatura (quatro), é suficiente para conseguir uma boa clusterização.

Como limitações, ainda, é importante ressaltar que, por insuficiência de recursos computacionais, não foi feita uma investigação mais ampla, considerando diferentes configurações iniciais para os algoritmos de clusterização e de classificação empregados, maior tempo usado em cada gravação EEG e outros algoritmos e conjuntos de dados disponíveis.

Dessa forma, como trabalhos futuros, além de superar as limitações descritas acima, sugere-se:

1. otimizar o hiperparâmetro K (número de microestados) frente a vários algoritmos de clusterização;
2. propor um método de clusterização otimizado usando conceitos presentes no estado da arte, como *agents* e *neurosymbolic*;
3. disponibilizar uma plataforma que possibilite à comunidade médica submeter os exames de imagem de forma segura e anonimizada, produzindo relatórios detalhados sobre os microestados, visualizando-os de forma tridimensional e comparando com o que se encontra na literatura de microestados para uma pessoa saudável.

REFERÊNCIAS

AL-KADI, M. I.; REAZ, M. B. I.; ALI, M. A. Evolution of electroencephalogram signal analysis techniques during anesthesia. **Sensors (Basel)**, Multidisciplinary Digital Publishing Institute, v. 13, n. 5, p. 6605–6635, 2013.

ANTONOVA, E.; HOLDING, M.; SUEN, H. C.; SUMICH, A.; MAEX, R.; NEHANIV, C. Eeg microstates: Functional significance and short-term test-retest reliability. **Neuroimage: Reports**, v. 2, n. 2, p. 100089, 2022. ISSN 2666-9560. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666956022000137>.

ANTONY, M. J.; SANKARALINGAM, B. P.; MAHENDRAN, R. K.; GARDEZI, A. A.; SHAFIQ, M.; CHOI, J.-G.; HAMAM, H. Classification of eeg using adaptive svm classifier with csp and online recursive independent component analysis. **Sensors**, v. 22, n. 19, 2022.

ARORA, P.; DEEPALI; VARSHNEY, S. Analysis of k-means and k-medoids algorithm for big data. **Procedia Computer Science**, v. 78, p. 507–512, 2016. ISSN 1877-0509. 1st International Conference on Information Security Privacy 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050916000971>.

ATTAR, E. Review of electroencephalography signals approaches for mental stress assessment. **Neurosciences (Riyadh)**, Riyadh: Saudi Neuroscience Society, v. 27, n. 4, p. 209–215, Oct 2022.

BANDITWATTANAWONG, T.; MASDISORNCHOTE, M. On characterization of norm-referenced achievement grading schemes toward explainability and selectability. **Applied Computational Intelligence and Soft Computing**, v. 2021, n. 1, p. 8899649, 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/8899649>.

Ben Ncir, C.-E.; HAMZA, A.; BOUAGUEL, W. Parallel and scalable dunn index for the validation of big data clusters. **Parallel Computing**, v. 102, p. 102751, 2021. ISSN 0167-8191. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167819121000119>.

BVSMS. **Aspectos gerais do avanço de Alzheimer no Brasil**. 2023. Aspectos gerais do avanço de Alzheimer no Brasil. Disponível em: <https://bvsms.saude.gov.br/nunca-e-cedo-demais-nunca-e-tarde-demais-setembro-mes-mundial-do-alzheimer/#:~:text=Segundo%20o%20relat%C3%B3rio%20mundial%20de,os%20maiores%20n%C3%BAmeros%20de%20casos.> Acesso em: 15 jun. 2024.

BVSMS. **O avanço dos casos de demência no Brasil**. 2023. O avanço dos casos de demência no Brasil. Disponível em: <https://bvsms.saude.gov.br/o-avanco-dos-casos-de-demencia-no-brasil-e-destaque-da-revista-pesquisa-fapesp-de-julho/>. Acesso em: 15 jun. 2024.

CHEN, W.; WANG, Y.; REN, Y.; JIANG, H.; DU, G.; ZHANG, J.; LI, J. An automated detection of epileptic seizures eeg using cnn classifier based on feature fusion with high accuracy. **BMC Med Inform Decis Mak**, v. 23, n. 1, p. 96, 2023.

CUENOUD, B.; IPEK, ; SHEVLYAKOVA, M.; BEAUMONT, M.; CUNNANE, S. C.; GRUETTER, R.; XIN, L. Brain nad is associated with atp energy production and membrane phospholipid turnover in humans. **Frontiers in Aging Neuroscience**, v. 12, 2020. ISSN 1663-4365. Disponível em: <https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2020.609517>.

CUI, R.; JIANG, J.; ZENG, L.; JIANG, L.; XIA, Z.; DONG, L.; GONG, D.; YAN, G.; MA, W.; YAO, D. Action video gaming experience related to altered resting-state eeg temporal and spatial complexity. **Frontiers in Human Neuroscience**, v. 15, 2021. ISSN 1662-5161. Disponível em: <https://www.frontiersin.org/articles/10.3389/fnhum.2021.640329>.

DU, Y.; LI, G.; WU, M.; CHEN, F. Unsupervised multivariate feature-based adaptive clustering analysis of epileptic eeg signals. **Brain Sciences**, v. 14, n. 4, 2024. ISSN 2076-3425. Disponível em: <https://www.mdpi.com/2076-3425/14/4/342>.

FATIMA, K.; MEHENDALE, A.; REDDY, H. Young-onset dementia and neurodegenerative disorders of the young with an emphasis on clinical manifestations. **Cureus**, Cureus, Inc., v. 14, n. 10, p. e30025, 2022.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. [S.l.]: Springer Science & Business Media, 2009.

HEARST, M.; DUMAIS, S.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. **IEEE Intelligent Systems and their Applications**, v. 13, n. 4, p. 18–28, 1998.

HUSSAIN, I.; JANY, R.; BOYER, R.; AZAD, A.; ALYAMI, S. A.; PARK, S. J.; HASAN, M. M.; HOSSAIN, M. A. An explainable eeg-based human activity recognition model using machine-learning approach and lime. **Sensors**, v. 23, n. 17, 2023. Disponível em: <https://www.mdpi.com/1424-8220/23/17/7452>.

IBGE. **Análise sobre a variação da expectativa de vida ao nascer com o passar dos anos**. 2023. Análise do panorama de expectativa de vida de 1940 a 2022. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/38455-em-2022-expectativa-de-vida-era-de-75-5-anos#:~:text=Na%20tabela%20a%20seguir%2C%20um,7%20anos%20para%20as%20mulheres.> Acesso em: 15 jun. 2024.

IDRIS, Z.; ZAKARIA, Z.; YEE, A. S.; FITZROL, D. N.; ISMAIL, M. I.; GHANI, A. R. I.; ABDULLAH, J. M.; HASSAN, M. H.; SUARDI, N. Light and the brain: A clinical case depicting the effects of light on brainwaves and possible presence of plasma-like brain energy. **Brain Sciences**, v. 14, n. 4, 2024. ISSN 2076-3425. Disponível em: <https://www.mdpi.com/2076-3425/14/4/308>.

IFTIMOVICI, A.; MARCHI, A.; FÉRAT, V.; PRUVOST-ROBIEUX, E.; GUINARD, E.; MORIN, V.; ELANDALOUSSI, Y.; D'HALLUIN, A.; KREBS, M.; CHAUMETTE, B.; GAVARET, M. Electroencephalography microstates imbalance across the spectrum of early psychosis, autism, and mood disorders. **European Psychiatry**, Cambridge University Press, v. 66, n. 1, p. e41, 2023.

IKOTUN, A. M.; EZUGWU, A. E.; ABUALIGAH, L.; ABUHAIJA, B.; HEMING, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. **Information Sciences**, v. 622, p. 178–210, 2023. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025522014633>.

JIANG, X.; BIAN, G. B.; TIAN, Z. Removal of artifacts from eeg signals: A review. **Sensors**, v. 19, n. 5, p. 987, February 2019.

JOSHI, S. R.; HEADLEY, D. B.; HO, K. C.; PARé, D.; NAIR, S. S. Classification of brainwaves using convolutional neural network. In: **Proc Eur Signal Process Conf EUSIPCO**. [s.n.], 2019. p. 10.23919/eusipco.2019.8902952. Disponível em: <https://doi.org/10.23919/eusipco.2019.8902952>.

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

KIESSNER, A.-K.; SCHIRRMEISTER, R. T.; GEMEIN, L. A.; BOEDECKER, J.; BALL, T. An extended clinical eeg dataset with 15,300 automatically labelled recordings for pathology decoding. **NeuroImage: Clinical**, v. 39, p. 103482, 2023. ISSN 2213-1582. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2213158223001730>.

KLEM, G. H.; LüDERS, H. O.; JASPER, H. H.; ELGER, C. The ten-twenty electrode system of the international federation. **Electroencephalography and Clinical Neurophysiology**, v. 52, p. 3–6, 1999.

KUI, M.; XU, Y.; WANG, J.; CHENG, F. Research on the adaptability of typical denoising algorithms based on icesat-2 data. **Remote Sensing**, v. 15, n. 15, 2023. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/15/15/3884>.

LANCET. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. GBD 2021 Nervous System Disorders Collaborators, 2024.

LASSI, M.; FABBIANI, C.; MAZZEO, S.; BURALI, R.; VERGANI, A. A.; GIACOMUCCI, G.; MOSCHINI, V.; MORINELLI, C.; EMILIANI, F.; SCARPINO, M.; BAGNOLI, S.; INGANNATO, A.; NACMIAS, B.; PADIGLIONI, S.; MICERA, S.; SORBI, S.; GRIPPO, A.; BESSI, V.; MAZZONI, A. Degradation of eeg microstates patterns in subjective cognitive decline and mild cognitive impairment: Early biomarkers along the alzheimer's disease continuum? **NeuroImage: Clinical**, v. 38, p. 103407, 2023. ISSN 2213-1582. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2213158223000967>.

LEHMANN, D.; OZAKI, H.; PAL, I. Eeg alpha map series: brain micro-states by space-oriented adaptive segmentation. **Electroencephalography and Clinical Neurophysiology**, v. 67, n. 3, p. 271–288, 1987. ISSN 0013-4694. Disponível em: <https://www.sciencedirect.com/science/article/pii/0013469487900253>.

LI, J.; HASSAN, D.; BREWER, S.; SITZENFREI, R. Is clustering time-series water depth useful? an exploratory study for flooding detection in urban drainage systems. **Water**, v. 12, n. 9, 2020. ISSN 2073-4441. Disponível em: <https://www.mdpi.com/2073-4441/12/9/2433>.

LIGHT, G. A.; WILLIAMS, L. E.; MINOW, F.; SPROCK, J.; RISSLING, A.; SHARP, R.; SWERDLOW, N. R.; BRAFF, D. L. Electroencephalography (eeg) and event-related potentials (erps) with human participants. **Current Protocols in Neuroscience**, Chapter 6, p. Unit 6.25.1–24, Jul 2010.

LIU, E.; LUU, C.; WU, L. C. Resting state eeg variability and implications for interpreting clinical effect sizes. **IEEE Transactions on Neural Systems and Rehabilitation Engineering**, v. 32, p. 587–596, 2024.

LIU, H.; ZHANG, Y.; LI, Y.; KONG, X. Review on emotion recognition based on electroencephalography. **Frontiers in Computational Neuroscience**, v. 15, 2021. ISSN 1662-5188. Disponível em: <https://www.frontiersin.org/articles/10.3389/fncom.2021.758212>.

LóPEZ, S.; SUAREZ, G.; JUNGREIS, D.; OBEID, I.; PICONE, J. Automated identification of abnormal adult eegs. **IEEE Signal Processing in Medicine and Biology Symposium (SPMB)**, p. 10.1109/SPMB.2015.7405423, 2015.

MA, Q.; WANG, M.; HU, L.; ZHANG, L.; HUA, Z. A novel recurrent neural network to classify eeg signals for customers' decision-making behavior prediction in brand extension scenario. **Frontiers in Human Neuroscience**, v. 15, 2021.

MACKINTOSH, A.; BORGWARDT, S.; STUDERUS, E.; RIECHER-RÖSSLER, A.; BOCK, R. de; ANDREOU, C. Eeg microstate differences in medicated vs. medication-naïve first-episode psychosis patients. **Frontiers in Psychiatry**, Frontiers Media SA, v. 11, p. 600606, 2020.

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

MURPHY, M.; STICKGOLD, R.; PARR, M. *et al.* Recurrence of task-related electroencephalographic activity during post-training quiet rest and sleep. **Scientific Reports**, v. 8, p. 5398, 2018. Disponível em: <https://doi.org/10.1038/s41598-018-23590-1>.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, v. 7, 2013. ISSN 1662-5218. Disponível em: <https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021>.

NUNES, T. M.; COELHO, A. L.; LIMA, C. A.; PAPA, J. P.; de Albuquerque, V. H. C. Eeg signal classification for epilepsy diagnosis via optimum path forest – a systematic assessment. **Neurocomputing**, v. 136, 2014.

OBAYYA, M.; SAEED, M. K.; MAASHI, M.; ALOTAIBI, S. S.; SALAMA, A. S.; HAMZA, M. A. A novel automated parkinson's disease identification approach using deep learning and eeg. **PeerJ Comput Sci**, v. 9, p. e1663, Nov 2023.

OBEID, I.; PICONE, J. The temple university hospital eeg data corpus. **Frontiers in Neuroscience**, Frontiers, v. 10, p. 196, 2016.

OMS. **Plano de ação global intersetorial sobre epilepsia e outras doenças neurológicas 2022–2031 (IGAP)**. 2022. Plano de ação global intersetorial sobre epilepsia e outras doenças neurológicas 2022–2031 (IGAP) DA OMS. Disponível em: <https://www.who.int/publications/i/item/9789240076624>. Acesso em: 15 jun. 2024.

PAHO. **The burden of Neurological conditions in the Region of the Americas, 2000-2019**. 2021. Estudo e visualização interativa do avanço de doenças neurológicas para países das Américas feito pela Organização Pan-Americana de Saúde. Disponível em: <https://www.paho.org/en/enlace/burden-neurological-conditions>. Acesso em: 15 jun. 2024.

PANDE, V. S.; BEAUCHAMP, K.; BOWMAN, G. R. Everything you wanted to know about markov state models but were afraid to ask. **Methods**, v. 52, n. 1, p. 99–105, Sep 2010.

PASCUAL-MARQUI, R. D.; MICHEL, C. M.; LEHMANN, D. Segmentation of brain electrical activity into microstates: model estimation and validation. **IEEE Transactions on Biomedical Engineering**, v. 42, n. 7, p. 658–665, July 1995.

PATEL, V. R.; MEHTA, R. G. Modified k-means clustering algorithm. In: DAS, V. V.; THANKACHAN, N. (Ed.). **Computational Intelligence and Information Technology**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 307–312. ISBN 978-3-642-25734-6.

PINEGGER, A.; WRIESSNEGGER, S. C.; FALLER, J.; MüLLER-PUTZ, G. R. Evaluation of different eeg acquisition systems concerning their suitability for building a brain–computer interface: Case studies. **Frontiers in Neuroscience**, v. 10, 2016. ISSN 1662-453X. Disponível em: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00441>.

POIAN, A. T. D.; EL-BACHA, T.; LUZ, M. R. M. P. Nutrient utilization in humans: Metabolism pathways. **Nature Education**, Nature Publishing Group, v. 3, n. 9, p. 11, 2010. Instituto de Bioquimica Medica, Universidade Federal do Rio de Janeiro; Instituto Oswaldo Cruz, Fundacao Oswaldo Cruz.

QUYEN, M. L. V.; MARTINERIE, J.; NAVARRO, V.; BOON, P.; D'HAVé, M.; ADAM, C.; RENAULT, B.; VARELA, F.; BAULAC, M. Anticipation of epileptic seizures from standard eeg recordings. **Lancet**, v. 357, n. 9251, p. 183–188, Jan 2001.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.

S, D. S. L. Reducing power line noise in eeg and meg data via spectrum interpolation. **Cureus**, v. 189, p. 763–776, 2019.

SCHILLING, M. W.; COGGINS, P. C. Utilization of agglomerative hierarchical clustering in the analysis of hedonic scaled consumer acceptability data. **Journal of Sensory Studies**, v. 22, p. 477–491, 2007. Disponível em: <https://doi.org/10.1111/j.1745-459X.2007.00121.x>.

SCHLEGEL, F.; LEHMANN, D.; FABER, P. L.; MILZ, P.; GIANOTTI, L. R. Eeg microstates during resting represent personality differences. **Brain Topography**, Springer, v. 25, n. 1, p. 20–26, 2012.

SETIAWAN, A. F.; WIBAWA, A. D.; PURNOMO, M. H.; ISLAMIYAH, W. R. Monitoring stroke rehabilitation re-learning program using eeg parameter: A preliminary study for developing self-monitoring system for stroke rehabilitation during new normal. In: **2020 International Seminar on Application for Technology of Information and Communication (iSemantic)**. [S.l.: s.n.], 2020. p. 620–624.

SONG, Y.; LU, Y. Decision tree methods: Applications for classification and prediction. **Shanghai Arch Psychiatry**, v. 27, n. 2, p. 130–135, Apr 25 2015.

STEVENS, A.; KIRCHER, T. Cognitive decline unlike normal aging is associated with alterations of eeg temporo-spatial characteristics. **European Archives of Psychiatry and Clinical Neuroscience**, Springer, v. 248, n. 5, p. 259–266, 1998.

SUN, C.; MOU, C. Survey on the research direction of eeg-based signal processing. **Frontiers in Neuroscience**, v. 17, 2023.

TARAILIS, P.; KOENIG, T.; MICHEL, C. *et al.* The functional aspects of resting eeg microstates: A systematic review. **Brain Topogr**, v. 37, p. 181–217, 2024. Disponível em: <https://doi-org.ez11.periodicos.capes.gov.br/10.1007/s10548-023-00958-9>.

TARWIDI, D.; PUDJAPRASETYA, S. R.; ADYTIA, D.; APRI, M. An optimized xgboost-based machine learning method for predicting wave run-up on a sloping beach. **MethodsX**, v. 10, p. 102119, 2023. ISSN 2215-0161. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2215016123001206>.

TOIVONEN, L.; FORSSTRöM, V.; WARIS, M.; PELTOLA, V. Acute respiratory infections in early childhood and risk of asthma at age 7 years. **Journal of Allergy and Clinical Immunology**, v. 143, n. 1, p. 407–410.e6, 2019. ISSN 0091-6749. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0091674918312697>.

TOMESCU, M. I.; PAPASTERI, C. C.; SOFONEA, A.; BOLDASU, R.; KEBETS, V.; PISTOL, C. A.; POALELUNGI, C.; BENESCU, V.; PODINA, I. R.; NEDELCEA, C. I.; BERCEANU, A. I.; CARCEA, I. Spontaneous thought and microstate activity modulation by social imitation. **NeuroImage**, v. 249, p. 118878, 2022. ISSN 1053-8119. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1053811922000088>.

TUDOR, M.; TUDOR, L.; TUDOR, K. I. Hans berger (1873-1941)–the history of electroencephalography. **Acta Med Croatica**, v. 59, n. 4, p. 307–313, 2005.

UDDIN, S.; HAQUE, I.; LU, H.; AL. et. Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction. **Sci Rep**, v. 12, p. 6256, 2022. Disponível em: <https://doi.org/10.1038/s41598-022-10358-x>.

WEGNER, F. von; KNAUT, P.; LAUFS, H. Eeg microstate sequences from different clustering algorithms are information-theoretically invariant. **Frontiers in Computational Neuroscience**, v. 12, 2018. ISSN 1662-5188. Disponível em: <https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2018.00070>.

WEN, T. Y.; ARIS, S. A. M. Hybrid approach of eeg stress level classification using k-means clustering and support vector machine. **IEEE Access**, v. 10, p. 18370–18379, 2022.

XAVIER, G.; TING, A. S.; FAUZAN, N. Exploratory study of brain waves and corresponding brain regions of fatigue on-call doctors using quantitative electroencephalogram. **J Occup Health**, v. 62, n. 1, p. e12121, Jan 2020.

YU, W.-Y.; SUN, T.-H.; HSU, K.-C.; WANG, C.-C.; CHIEN, S.-Y.; TSAI, C.-H.; YANG, Y.-W. Comparative analysis of machine learning algorithms for alzheimer's disease classification using eeg signals and genetic information. **Computers in Biology and Medicine**, v. 176, 2024.

ZIEGLER, A.; KöNIG, I. R. Mining data with random forests: Current options for real-world applications. **WIREs Data Mining and Knowledge Discovery**, v. 4, p. 55–63, 2014. Disponível em: <https://doi.org/10.1002/widm.1114>.

ZOUBI, O. A.; MAYELI, A.; TSUCHIYAGAITO, A.; MISAKI, M.; ZOTEV, V.; REFAI, H.; PAULUS, M.; BODURKA, J.; , t. T. . I. ; AUPPERLE, R. L.; KHALSA, S. S.; FEINSTEIN, J. S.; SAVITZ, J.; CHA, Y.-H.; KUPLICKI, R.; VICTOR, T. A. Eeg microstates temporal dynamics differentiate individuals with mood and anxiety disorders from healthy subjects. **Frontiers in Human Neuroscience**, v. 13, 2019. ISSN 1662-5161. Disponível em: <https://www.frontiersin.org/articles/10.3389/fnhum.2019.00056>.