תמונה שמכילה גופן, טקסט, לוגו, גרפיקה

התיאור נוצר באופן אוטומטי

Capstone Project Phase B

**Classification of ADHD vs. Healthy Children Using Pattern Matching on EEG Signals**

24-1-R-5

Mohammed Khatib, Basel Kabia

Supervisors: Dr. Anat Dahan, Dr. Samah Idrees-Ghazawi

**Table of Contents**

1 Introduction. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . 3

2 Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . 4

2.1 Preprocessing and Data Cleaning. . . . . . . . . . . . .. . . . . 4

2.2 First approach using BLAST . . .. . . . . . . . .. . . . . .. . ... 6

2.3 Second approach using order-preserving matching.. . . . 7

2.4 Classification process . . . . . . . . . . . . . . . . . . . . . . . . . . . .9

3 **Challenges and Solutions**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. 9

4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …. .. ….10

4.1 BLAST Result . . . . . . . . . . . . . . . . . . . . . . . . . . …. ... ….10

4.2 Order-Preserving Matching Result . . . . . . . . . . . . . . ... 14

5 **Conclusions**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …………... 15

5.1 **Achievement of Project Goals**. . . . . . . . . . . . . . . . . . . . . **15**

5.2 **Handling Challenges and Decision-Making** . . . . . . . . . . .**15**

6 Project Reflections. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **15**

**6.1 Evaluation of Approach and Potential Changes** . . . . . .. ..15

6.2 **Meeting Project Metrics** . . . . . . . . . . . . . . . . . . . . . . . . . . 15

**Abstract.** This study aims to enhance EEG classification for distinguishing between ADHD and Control subjects using two analytical approaches: BLAST comparisons and order-preserving matching. We investigate the effectiveness of various time segments (1 to 15 seconds) and alignment methods with BLAST to identify the optimal segment and channel combinations for accurate classification. Additionally, we explore order-preserving matching to evaluate its potential as an alternative method. Our goal is to determine the best time segment and channel for high classification accuracy while balancing computational efficiency, and to assess the practicality of order-preserving matching for EEG analysis.

**Keywords**: Electroencephalography (EEG), Attention-Deficit/Hyperactivity Disorder (ADHD), Pattern matching algorithms, Computational biology, BLAST, Order-preserving matching.

**1. Introduction**

Electroencephalography (EEG) plays a pivotal role in diagnosing and understanding neurological conditions by capturing the brain's electrical activity. Accurate classification of EEG data is essential for distinguishing between conditions such as Attention Deficit Hyperactivity Disorder (ADHD) and Control groups. This project aims to enhance the classification process by evaluating and comparing two distinct analytical approaches: BLAST comparisons and order-preserving matching.

The primary objectives of this study were to identify the most effective EEG channels for classifying ADHD and Control subjects, assess the performance of different alignment methods using BLAST, and analyze various time segments to determine the optimal length for classification accuracy and processing efficiency. Specifically, the project sought to determine which time segment and channel combinations would provide the highest accuracy in distinguishing between ADHD and Control groups. Additionally, the study aimed to explore the potential of order-preserving matching as an alternative method for classifying EEG data, offering a different perspective on the classification process.

The significance of this study lies in its potential to refine diagnostic tools and improve understanding of neurological conditions through better classification techniques. By comparing different time segments and analytical methods, the study provides insights into achieving a balance between accuracy and computational efficiency. This is crucial for developing more effective and practical diagnostic practices. The project addresses key computational challenges and aims to guide future research and applications in the field of EEG analysis.

The methodologies employed in this project included BLAST comparisons, which involved analyzing EEG data using longest alignments and highest percentage identity alignments across various time segments. This approach was designed to identify the most effective channels and time segments for classification. Additionally, the study explored order-preserving matching, which maintains the sequence order of the EEG signals, offering an alternative approach to classification. Through these methodologies, the project aims to provide a comprehensive understanding of the optimal conditions for EEG classification and offer valuable insights for future research and clinical applications.

**2. Experiments**

In this study, after cleaning the data, we will employ two different approaches: one using BLAST (one by the longest alignment and the other by the highest alignment) and the other using order-preserving matching. Subsequently, we will classify the results based on these methods (Figure 1).

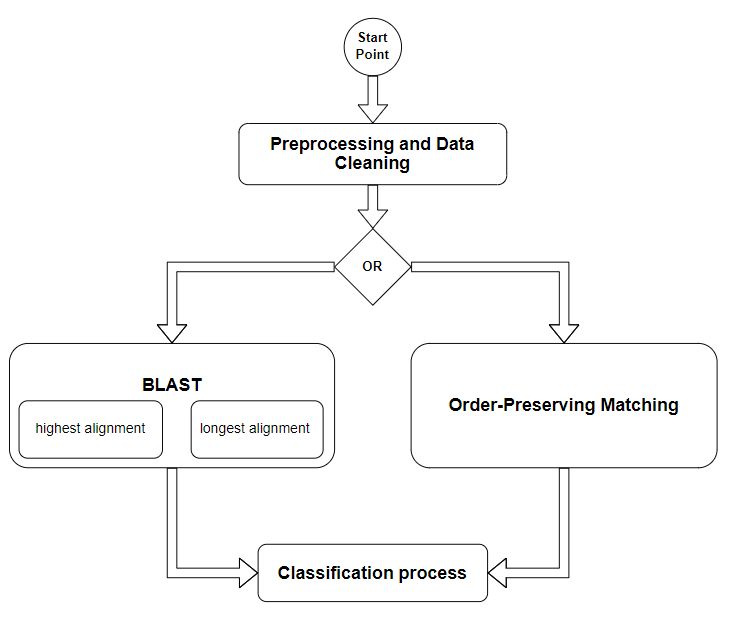


Fig. 1: flowchart that describe study steps.

**2.1 Preprocessing and Data Cleaning**

To ensure the EEG data is of high quality and free from artifacts, a comprehensive preprocessing and data cleaning procedure was implemented. The raw EEG data, stored in `.mat` files, was first loaded using the `scipy.io` library. The specific variable containing the EEG data was extracted, and the data was configured with a sampling frequency of 128 Hz. The EEG channels were labeled according to the standard 10-20 electrode placement system, including channels such as Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, and Pz (Figure 2). This labeling is crucial for ensuring that the data is accurately aligned with standardized electrode positions.

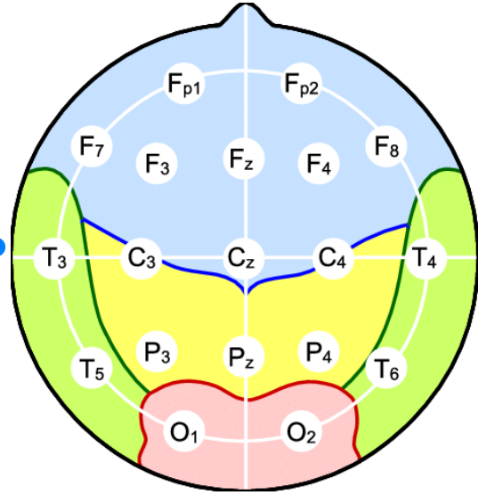
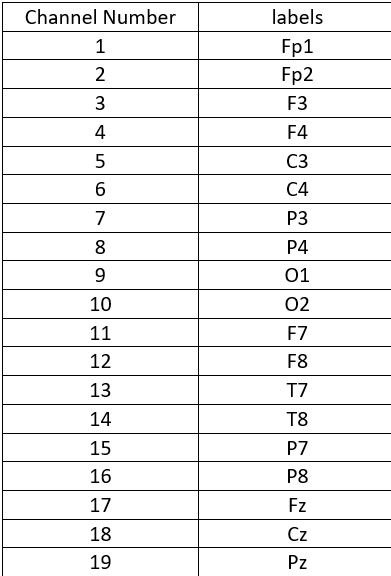
 

Fig. 2: This figure illustrates the channel locations in the 10-20 system of EEG.

An MNE `RawArray` object was then created using the loaded data and channel information, which serves as a foundational structure for further processing. To enhance the quality of the EEG signals, a standard 10-20 electrode placement montage was applied to the data. Following this, the EEG data was band-pass filtered between 0.5 and 40 Hz using a FIR filter design. This filtering step is essential for removing low-frequency drift and high-frequency noise, which could otherwise interfere with signal interpretation and analysis.

To address common EEG artifacts, such as those caused by eye blinks or muscle activity, Independent Component Analysis (ICA) was applied. The ICA procedure was set to extract 19 components, corresponding to the number of EEG channels. After fitting the ICA model to the data, components were manually inspected, and specific components were excluded based on their contribution to artifacts. In this example, components 0 and 1 were excluded, though this step would typically involve more detailed visual inspection.

The cleaned EEG data was then saved back to `.mat` files for further analysis. To maintain organization, separate directories were used for the ADHD and control groups, ensuring that the data remained clearly categorized. Finally, the entire cleaning process was automated to handle multiple files within the ADHD and control directories. The code iterated through all `.mat` files, applied the cleaning procedure, and saved the cleaned data to the appropriate directories, streamlining the preprocessing phase of the study.

**2.2 First approach using BLAST**

BLAST, or Basic Local Alignment Search Tool, is a widely used algorithm for comparing sequences of DNA, RNA, or proteins. It helps identify regions of similarity between sequences by finding local matches, making it useful for discovering relationships between genes, proteins, or other biological data. Instead of aligning entire sequences from end to end, BLAST focuses on matching the most similar parts, which makes it faster and more efficient, especially when searching large databases. Researchers use BLAST to study the function and evolution of genes, identify unknown sequences, and make sense of large genetic datasets by comparing them to known sequences.

The conversion of EEG data into letter sequences is a critical step in our analysis, aiming to symbolize brain wave activity over time. This process begins with identifying the most dominant frequency within each segment of the EEG signal, as this frequency reflects the primary brain wave activity during that time. To achieve precise frequency analysis, the EEG signal is divided into small, overlapping time windows, allowing for a detailed frequency spectrum over time.

Formula for Power Spectral Density (PSD):

1. **EEG Signal**: Let represent the EEG signal in the time domain.
2. **Fourier Transform**: First, compute the Fourier Transform of to obtain the frequency domain representation
3. **Power Spectrum**: The power spectrum is calculated as the squared magnitude of the Fourier Transform:
4. **Power Spectral Density (PSD)**: PSD is typically normalized to account for the signal length or sampling rate, so the formula for the PSD can be given as: ​ where T is the length of the signal (or the number of samples if you're working in discrete time).

The PSD reveals how power is distributed across various frequencies, and the frequency with the highest power is identified as the dominant frequency for that segment.

This dominant frequency is then mapped to a specific letter to create a symbolic representation of the EEG data. For example, low frequencies such as Delta waves (0.5-4 Hz) are mapped to the letter 'A,' while higher frequencies such as Gamma waves (30-40 Hz) are mapped to the letter 'U.' This mapping transforms continuous EEG data into letter sequences, each representing dominant brain wave activity at different time points.

The primary purpose of this transformation is to enable symbolic comparison using bioinformatics tools like BLAST (Basic Local Alignment Search Tool). By converting EEG data into sequences of letters, these sequences can be compared across different subjects or conditions in a manner similar to genetic sequence comparison in genomics. This approach allows for the identification of patterns or similarities in brain wave activity, potentially revealing specific brain states or disorders.

BLAST is a tool traditionally used in bioinformatics, was adapted in this study to compare EEG-derived letter sequences. These sequences, representing dominant EEG frequencies, were compared across ADHD and control groups to identify similarities or differences in brain wave patterns.

The process begins by generating sequences from both groups, with each sequence formatted in FASTA style, commonly used for biological sequences. The sequences from ADHD subjects serve as the "query," while those from control subjects form the "database" for comparison. BLAST is then employed to align these sequences, identifying regions of similarity between the EEG patterns of ADHD and control groups. Specific alignment parameters were selected to detect subtle but potentially meaningful similarities.

The BLAST results are analyzed to extract the longest alignments and those with the highest percentage identity across different EEG channels. The longest alignment indicates extended regions of similarity, while the highest identity percentage highlights the precision of the match between sequences. These results are then analyzed to identify consistent EEG pattern differences between ADHD and control groups, which may correlate with the studied condition.

This method leverages BLAST's ability to manage large datasets and uncover subtle similarities that might be missed by other approaches, providing valuable insights into EEG data and potential neurophysiological differences in ADHD.

**2.3 Second approach using Order-Preserving Matching**

In this study, EEG data was preprocessed by downsampling to reduce the data size while preserving the key features of the signals. Downsampling was performed by retaining every 4th sample from the original signal to reduce computational complexity without significantly impacting the integrity of the signal. This step is crucial to balance the memory usage and processing time with the need to retain essential temporal patterns.

The EEG data, extracted from .mat files, contains 19 channels per recording. Each channel's data is converted into a sequence of symbols representing positive (+) and negative (-) values. This symbolic representation allows for order-preserving matching, facilitating the comparison of patterns across different subjects.

For the pairwise comparison of EEG channels, we employed an order-preserving matching technique based on the **Needleman-Wunsch global alignment algorithm**. The Needleman-Wunsch algorithm is traditionally used for sequence alignment in bioinformatics, but its global alignment capabilities are leveraged here to compare entire EEG sequences. This method ensures that the temporal order of the data is preserved throughout the alignment process.

The algorithm aligns two sequences by maximizing matches and minimizing gaps (insertions/deletions) while preserving the order of elements. For each pair of EEG channels, the sequences are aligned, and the percentage similarity is calculated as follows:

The analysis included pairwise comparisons of EEG data across two groups: ADHD and Control. Each subject's EEG data was compared against others within the same group (e.g., ADHD-ADHD, Control-Control) as well as across groups (e.g., ADHD-Control). The order-preserving matching approach ensured that even subtle variations in the EEG signals were captured, allowing for a detailed comparison between different subjects and conditions.

For each comparison, the similarity percentages were calculated for all 19 EEG channels. The results were stored and analyzed to identify patterns that distinguish the ADHD group from the Control group.

To handle the computational load of comparing multiple EEG recordings, several optimization strategies were implemented:

1. **Downsampling**: Reduced the data points for each EEG channel to speed up processing while maintaining the integrity of the original signals.
2. **Efficient File Handling**: EEG data from each subject was stored in separate files, with the comparison results written to a unified output file, minimizing memory overhead.

These optimizations allowed for the efficient processing and comparison of large EEG datasets, enabling the identification of key differences in EEG patterns between ADHD and Control subjects.

**2.4 Classification process**

The classification process in this research is based on the analysis of the BLAST and order-preserving matching comparison results between the sequences derived from EEG data of ADHD and control subjects. The method focuses on identifying which group (ADHD or Control) each sequence more closely aligns with, using the two approaches identity percentages as a key metric.

The methods results are processed by reading the identity percentages of sequence alignments for each channel of the EEG data. The sequences are grouped based on whether they come from ADHD or Control subjects, and the average identity percentage is calculated for each channel across all comparisons. The classification is then made by comparing the average identity percentage of each sequence when aligned with ADHD sequences versus control sequences.

For each EEG channel, the program calculates whether the sequence from a given subject aligns more closely with the ADHD group or the control group. This is done by averaging the identity percentages of all alignments where the subject is compared to either group. If the average identity percentage is higher when compared to the ADHD group, the sequence is classified as ADHD for that channel; otherwise, it is classified as Control.

The classification results are tallied for each channel to determine how many sequences were correctly classified as ADHD or Control. The final output indicates the channel with the highest correct classifications for each group, providing insight into which EEG channels are most indicative of ADHD or Control patterns.

**3 Challenges and Solutions**

Throughout the study, we faced several analytical and technical challenges related to data processing, algorithm efficiency, and computational constraints. Below is an overview of these challenges and the strategies we implemented to overcome them.

**1. Handling BLAST Efficiency** Our initial approach involved utilizing BLAST with an online database to compare EEG-derived letter sequences. However, the process was prohibitively time-consuming due to the large data volume and real-time sequence comparison requirements. This led us to transition to BLAST+, a standalone version of BLAST that works with a local database. By using BLAST+ locally, we significantly reduced computation time and gained better control over alignment parameters, allowing for more efficient and customized analysis.

**2. Adapting to Specific Letter Sequences** BLAST is designed to work with biological sequences, such as DNA or protein sequences composed of specific letters. Adapting BLAST for EEG-derived letter sequences representing brain wave frequencies required thoughtful consideration of the letter mappings. We had to ensure that these mappings were consistent and meaningful in the context of brain wave patterns while maintaining compatibility with the BLAST tool. This required refining our mapping strategy to balance biological relevance with computational requirements.

**3. Global Alignment and Sequence Length** When employing global alignment methods, such as the Needleman-Wunsch algorithm, the length and complexity of the EEG-generated sequences became a challenge. Long sequences made the alignment process slow and difficult to interpret. To address this, we implemented downsampling techniques that reduced the resolution of the sequences while preserving key information. This improved the efficiency of the global alignment process, reducing computational load and enhancing result interpretability.

**4. Local Maxima and Minima Method** We also explored an alternative approach where we attempted to identify local maxima and minima in the raw EEG signal for each channel. The idea was to segment the data based on these local extrema and convert the segments into a 20-letter code, making it compatible with BLAST protein sequence alignment. However, this method proved impractical, as it required days of runtime due to the complexity of performing local alignment on the entire dataset. After assessing the computational feasibility, we decided to cancel this approach, focusing instead on methods that offered a better balance between performance and insight generation.

**4 Results**

**4.1 BLAST Result**

In the first approach, our analysis aimed to identify the most effective EEG channels for classifying ADHD and Control subjects by employing BLAST comparisons. We specifically examined the performance of the longest alignments and the highest percentage identity alignments across various time segments, namely 1, 2, 3, 5, 7, 9, 12, and 15 seconds. The objective was to determine which time segment and channel combinations yield the highest classification accuracy between the ADHD and Control groups.

During the evaluation, we closely assessed both the processing time and classification accuracy for each time segment. Our findings indicated that the 1-second and 2-second segments produced nearly identical classification accuracy, with both segments yielding the best results. Specifically, for the 1-second time segment, we achieved a total of 110 out of 121 correct diagnoses using the highest percentage identity alignments, and 109 out of 121 correct diagnoses using the longest alignments. Similarly, the 2-second segment resulted in 109 out of 121 correct diagnoses for both alignment methods.

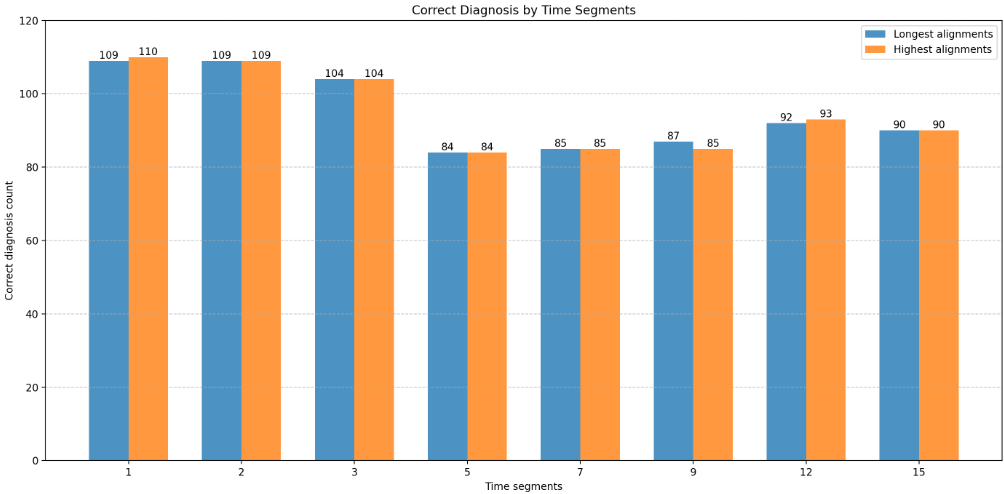
Despite the comparable accuracy of the 1-second and 2-second segments, the 2-second segment was found to be more advantageous overall. This is due to its significantly faster processing time of 16,526.59 seconds, in contrast to the 1-second segment, which required 77,677.26 seconds. Thus, the 2-second time segment offers a more practical balance between classification accuracy and computational efficiency, making it the preferred choice for achieving optimal performance in our analysis (refer to Figure 3 and Table 1 for a detailed investigation).

Fig. 3: chart that describe correct diagnosis count for each time segments in both longest and highest alignments.

|  |  |  |  |
| --- | --- | --- | --- |
| Time segments | Accuracy with longest alignments | Accuracy with highest alignments | Time (seconds) |
| 1 | 90% | 91% | 77677.26 |
| 2 | 90% | 90% | 16526.59 |
| 3 | 86% | 86% | 6164.66 |
| 5 | 69% | 69% | 2145.89 |
| 7 | 70% | 70% | 1227.81 |
| 9 | 72% | 70% | 883.76 |
| 12 | 76% | 77% | 807.21 |
| 15 | 74% | 74% | 420.62 |

Table 1: describe accuracy and time for each time segments in both longest and highest alignments.

The results presented in Figures 4, 5, 6, and 7 reveal that there is minimal difference between using BLAST for longest alignments versus highest percentage identity alignments in our analysis. Both methods yielded comparable performance across various time segments.

By focusing on the two most accurate segments, the 1-second and 2-second intervals, channel 16 emerged as the most effective for diagnosing ADHD. This suggests that channel 16 consistently achieved the highest classification accuracy for ADHD across these shorter time segments.

Conversely, the optimal channel for diagnosing Control subjects varied depending on the time segment. For the 1-second time segment, channel 19 proved to be the most accurate in diagnosing Control subjects. In contrast, for the 2-second time segment, channel 11 achieved the highest accuracy in classifying Control subjects.

These findings suggest that while the choice of alignment method has a negligible impact on classification performance, the effectiveness of individual channels for diagnosing ADHD and Control subjects can vary depending on the time segment used.

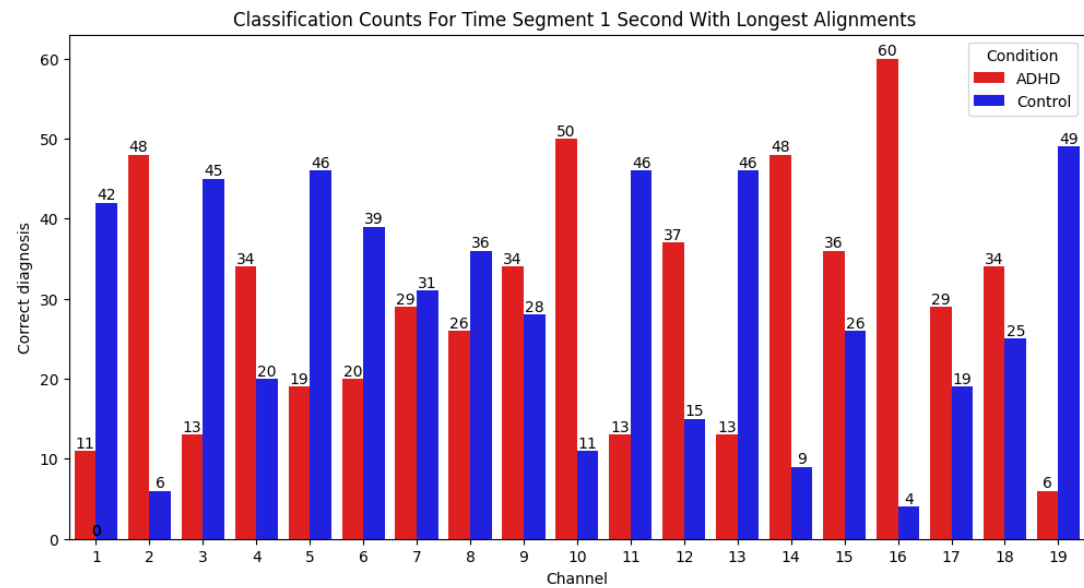
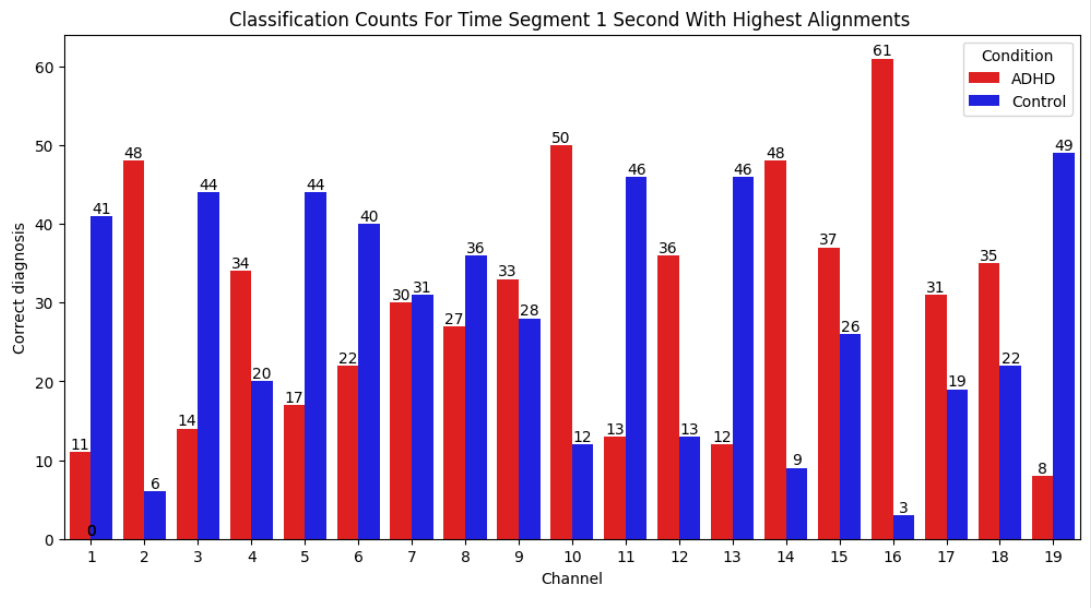




Fig. 4: chart that describe classification Counts by Channel for ADHD and Control in time segments of 1 second with longest alignment.

Fig. 5: chart that describe classification Counts by Channel for ADHD and Control in time segments of 1 second with highest alignment.

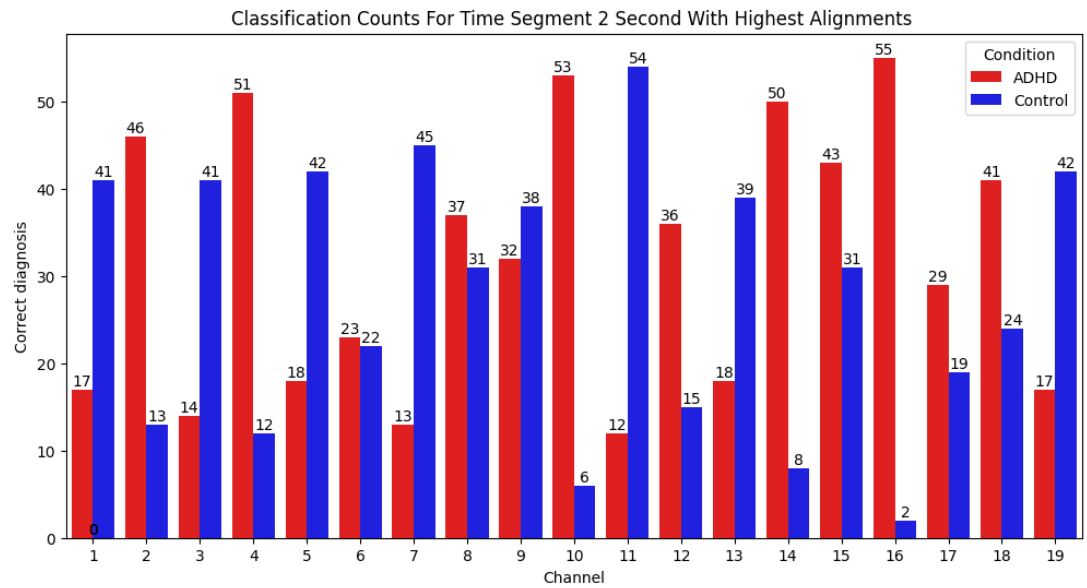


תמונה שמכילה טקסט, צילום מסך, עלילה, קו

התיאור נוצר באופן אוטומטי



Fig. 6: chart that describe classification Counts by Channel for ADHD and Control in time segments of 2 second with longest alignment.

 Fig. 7: chart that describe classification Counts by Channel for ADHD and Control in time segments of 2 second with highest alignment.



**4.2 Order-Preserving Matching** **Result**

In the second approach of our study, which employs order-preserving matching, the classification process was notably time-consuming, taking a total of 180,345.48 seconds, or approximately 50 hours. Despite this extensive processing time, the method achieved a correct classification rate for 87 out of 120 diagnoses. Specifically, channel 6 was utilized for identifying ADHD, while any channel could be used for classifying Control subjects.

It is important to highlight that, in this approach, nearly all channels were able to accurately classify the Control group, achieving a perfect classification rate of 59 out of 59 diagnoses for Controls. This indicates a high level of reliability and accuracy for identifying Control subjects across different channels. However, the extended duration required for the classification process underscores the need for optimizing computational efficiency in future analyses.

תמונה שמכילה טקסט, צילום מסך, מקביל, קו

התיאור נוצר באופן אוטומטי

Fig. 8: chart that describe correct diagnosis count for order-preserving matching.

**5 Conclusions**

**5.1 Achievement of Project Goals**

The project objectives were successfully achieved. We aimed to identify the most effective EEG channels for classifying ADHD and Control subjects and to evaluate the performance of different alignment methods and time segments. Using BLAST comparisons, we determined that the 2-second time segment provided the best balance between classification accuracy and processing efficiency. For ADHD, channel 16 consistently performed well, while channels 19 and 11 were most effective for Control diagnoses at 1-second and 2-second segments, respectively. The order-preserving matching approach, although highly accurate, highlighted the need for computational efficiency due to its lengthy processing time.

**5.2 Handling Challenges and Decision-Making**

We encountered challenges related to the computational demands of the order-preserving matching approach. To address this, we focused on optimizing the BLAST comparisons, which offered a more practical balance between accuracy and processing time. The decision to use a 2-second time segment was based on its superior efficiency and accuracy, making it a suitable choice for our objectives.

**6 Project Reflections**

**6.1 Evaluation of Approach and Potential Changes**

Overall, the methods employed were effective, but there are areas for improvement. In retrospect, optimizing computational efficiency for the order-preserving matching approach would be beneficial. This could involve exploring more efficient algorithms or hardware solutions to reduce processing time. Additionally, a deeper analysis of the impact of different time segments on classification accuracy could provide further insights for future projects.

**6.2 Meeting Project Metrics**

We met most of the project metrics we set for ourselves, including achieving high classification accuracy and identifying the optimal time segment and channel combinations. The choice of the 2-second segment with BLAST comparisons was particularly successful in balancing accuracy and efficiency. However, the extended processing time for the order-preserving matching approach highlighted the need for improved computational strategies, which we aim to address in future work.