Software Engineering Department

Braude College

Capstone Project Phase B

Order Pattern Matching for ADHD analysis using EEG waves

https://mega.nz/folder/nQxinQZA#YTR33_RAgW5PNK3yTuYVEQ

https://github.com/yaakovsh8/OPM4EEG.git

25-1-R-1

Lecturers Guide:
Dr. Anat Dahan
Dr. Samah Idrees Ghazawi

Students Submit:
Tamir Nahari
Yaakov Shitrit

Abstract

This study examined whether there were unique and recurring patterns of brain electrical activity among children aged 7–12 diagnosed with ADHD (Attention Deficit Hyperactivity Disorder) using EEG (Electroencephalography) technology. The EEG data were collected from 121 children, including 61 diagnosed with ADHD and 60 in a control group without psychiatric or neurological disorders. ADHD diagnoses were made by an experienced psychiatrist according to DSM-IV criteria, and all ADHD participants had been treated with Ritalin within six months before the study. EEG recordings were obtained from 19 scalp channels according to the 10–20 system at a sampling rate of 128 Hz, during a visual attention task in which children counted characters in rapidly presented cartoon images.

The research focused on identifying distinctive electrical activity patterns that differentiated children with ADHD from those without the disorder. Pattern matching techniques, specifically the Order Preserving Pattern Matching (OPM) algorithm, were applied to achieve this goal. The study aimed to determine the existence of these unique patterns, which could serve as a foundation for developing future tools for faster and more accurate ADHD diagnosis.

To identify ADHD patterns, we focused on two aspects (macro and micro).

Solution Approach 1: We examined whether the wave structure itself characterized a pattern that was the main difference between those with ADHD and those without ADHD.

Solution Approach 2: We investigated whether a specific wave sequence constituted a pattern that distinguished individuals with ADHD from those without ADHD.

A pattern was considered ADHD-related if it appeared in at least 90% of the ADHD group and rarely appeared in the control group.

Our final goal was to build a method that could support early ADHD detection and help doctors adjust treatments more personally. This kind of tool could make future diagnoses easier and faster.

1. Introduction

This study presents two approaches based on Order Preserving Matching (OPM) for identifying recurring EEG patterns among children aged 7–12 diagnosed with ADHD.

In the first approach (Approach 1), we apply the OPM algorithm to the structure of individual EEG waves—analyzing their morphology to uncover unique patterns frequently observed in ADHD subjects.

In the second approach (Approach 2), we focus on the sequence of wave occurrences: each wave is assigned a numerical value, and OPM is applied to the resulting sequence in order to detect patterns that differentiate the ADHD group from the control group.

The combination of these two approaches is intended to lay the foundation for a more accurate and efficient EEG-based diagnostic tool.

Currently, ADHD diagnostic tools serve as decision support systems rather than definitive diagnostic instruments—primarily due to the strict requirement of achieving over 90% accuracy in metrics such as PPV, NPV, Sensitivity, and Specificity.
For this reason, we based our core analysis on a didactic rather than probabilistic statistical approach, and framed the study around a fundamental research question:
Does the structure of an individual wave represent the pattern that distinguishes ADHD from typical development, or is it the sequence of wave occurrences that carries the critical distinction?

Nonetheless, for the purpose of EEG signal preprocessing, we utilized the AMICA algorithm—a probabilistic and advanced method for decomposing the EEG signal into independent components, aimed at enabling cleaner and more accurate analysis in subsequent stages.
We do not rely on the components themselves directly; rather, we select waveforms following decomposition and analyze them using the OPM algorithm.
The core analysis—searching for structural or sequential patterns—is not based on statistical inference, but rather on clear, consistent, and differentiable pattern occurrences predefined within the study groups.

This document (Part B of our research) presents the solution path chosen for each approach, in order to address this central research question.

Definition of Order Preserving Matching (OPM)

Order Preserving Matching (OPM) is an algorithm used to identify patterns whose **relative order** of values is preserved over time. Instead of comparing the exact numerical values of two sequences, OPM compares the **rank order** of their elements. Two sequences are considered a match if their relative ordering is the same, even if their actual amplitudes or numerical values differ.

In this project, we applied OPM by representing each EEG segment as a rank-order pattern, allowing us to detect recurring structures that maintained their shape despite variations in signal amplitude or scale. This property is particularly useful for EEG analysis, where brain activity may preserve a similar oscillatory structure even when absolute signal strength changes due to noise, physiological differences, or inter-subject variability.

2. PRE-PROCESSING

2.1 Preprocessing – Approach 1

In the initial phase of the study, a dedicated signal processing pipeline was developed to identify recurring waveform patterns in EEG recordings of children diagnosed with ADHD, in comparison to a control group of children without neurological or psychiatric diagnoses. This approach focused on the local structure of the EEG waveform—specifically the morphology of individual oscillatory cycles—in order to examine whether certain shapes occurred more frequently among subjects with ADHD.

The raw EEG signals first underwent standard preprocessing procedures, including high-pass and low-pass filtering, powerline notch filtering at 50 Hz, and Common Average Referencing (CAR). These steps were applied to reduce non-neural noise and eliminate baseline fluctuations that could obscure subtle waveform patterns.

Following preprocessing, the signals were decomposed into statistically independent components using the AMICA algorithm. The original compiled version of AMICA was executed via a custom wrapper developed as part of this study, ensuring precise control and stable execution. This decomposition yielded a set of independent time series (ICs), each assumed to reflect a distinct source of neural or artifactual activity.

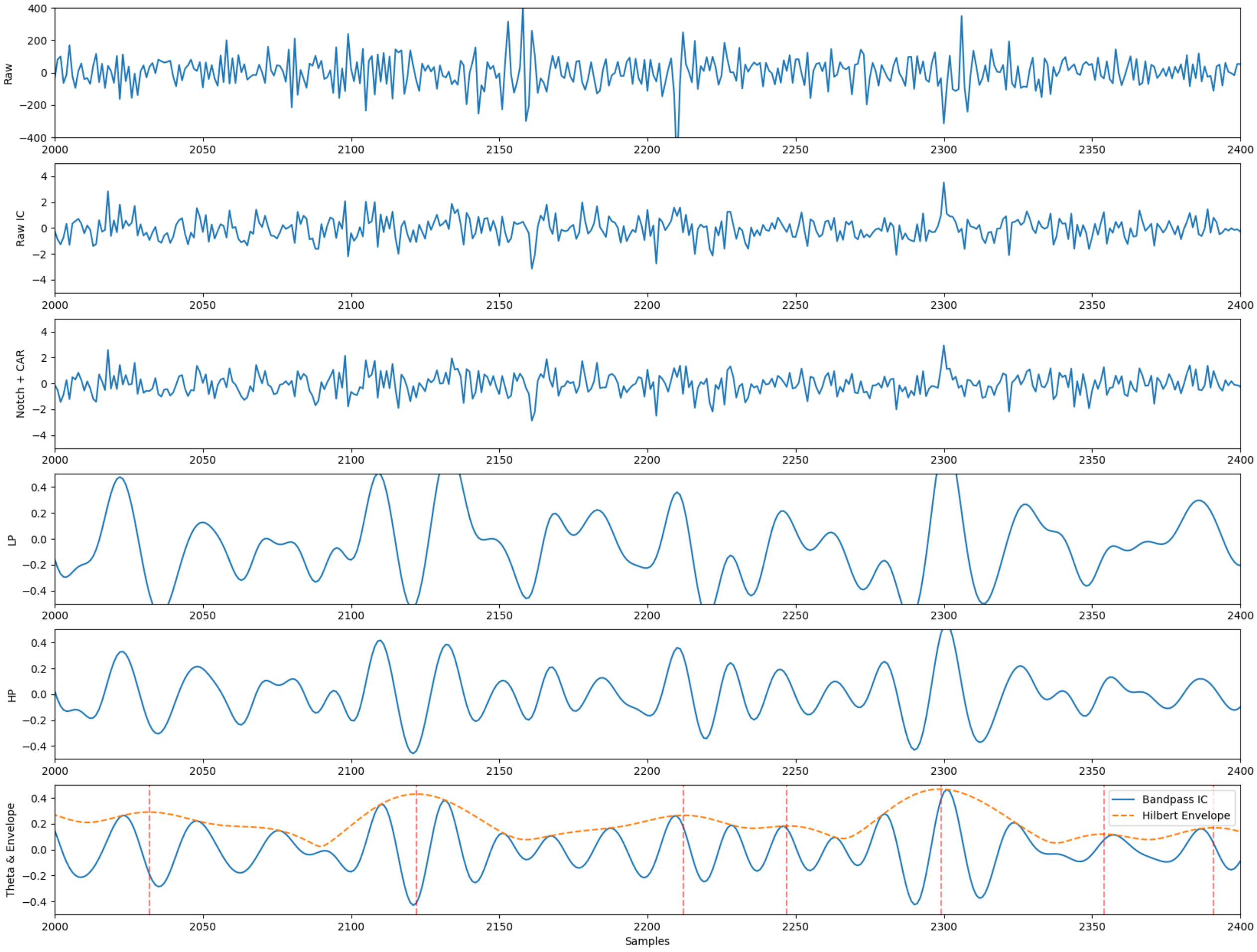
A band-pass filter in the 4–8 Hz theta range was then applied to each IC to isolate activity in a frequency band previously associated in the literature with attentional and frontal lobe processes, often implicated in ADHD. The filtered signals were then transformed using the Hilbert transform to extract both the amplitude envelope and the instantaneous phase.

Theta bursts were detected based on two criteria: the amplitude envelope had to exceed 1.5 standard deviations above the mean, and the cumulative phase progression during the burst had to reach at least 6π, corresponding to a minimum of three full oscillations. Within each burst, phase-crossing points were identified at 2π intervals, and a segment containing at least three full cycles was selected for further analysis.

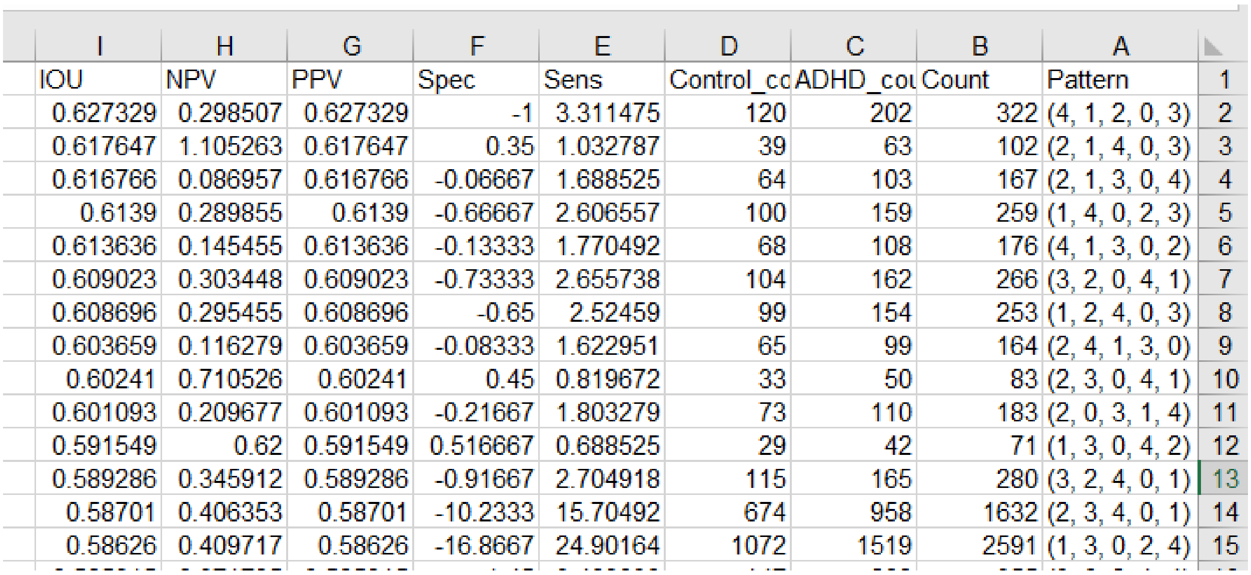
Each selected segment was then encoded using the Order Preserving Matching (OPM) method, which preserved the internal relative ranking of signal amplitudes over time. This symbolic transformation allowed the waveform shape to be captured in a compact, scale-invariant form. The resulting patterns were stored in a hash table to enable efficient pattern matching across other EEG segments.

To avoid the quadratic blow-up of pairwise pattern comparisons (which can take days on large datasets), we then group all OPM-encoded sequences into a hash table keyed by their OPM symbol series. Every waveform whose OPM sequence is identical ends up in the same bucket. Finally, we scan only those buckets whose size exceeds one; each such bucket contains at least two waveforms that match under the usual comparison metric. This simple hashing strategy reduces our entire pattern-matching step from days of computation down to mere minutes.

This approach is unique in the fact that it is based on a didactic approach that does not rely on statistical measures. Although the AMICA algorithm is used to clean the signal and isolate relevant waves, thereby making the decision which wave to examine using the OPM series, the decision as to whether it is a significant pattern is made solely according to the degree of correspondence of the series found to other series – that is, according to the OPM series itself. This means that the entire process is carried out on the basis of predefined rules, without the use of probabilistic tests, and therefore it is a didactic rather than statistical approach.



The result after analyzing the waves is the output of OPM patterns according to their frequency of occurrence among the ADHD group versus the control group and the calculations of the index values:



I	H	G	F	E	D	C	B	A	
IOU	NPV	PPV	Spec	Sens	Control_cc	ADHD_cou	Count	Pattern	1
0.627329	0.298507	0.627329	-1	3.311475	120	202	322	(4, 1, 2, 0, 3)	2
0.617647	1.105263	0.617647	0.35	1.032787	39	63	102	(2, 1, 4, 0, 3)	3
0.616766	0.086957	0.616766	-0.06667	1.688525	64	103	167	(2, 1, 3, 0, 4)	4
0.6139	0.289855	0.6139	-0.66667	2.606557	100	159	259	(1, 4, 0, 2, 3)	5
0.613636	0.145455	0.613636	-0.13333	1.770492	68	108	176	(4, 1, 3, 0, 2)	6
0.609023	0.303448	0.609023	-0.73333	2.655738	104	162	266	(3, 2, 0, 4, 1)	7
0.608696	0.295455	0.608696	-0.65	2.52459	99	154	253	(1, 2, 4, 0, 3)	8
0.603659	0.116279	0.603659	-0.08333	1.622951	65	99	164	(2, 4, 1, 3, 0)	9
0.60241	0.710526	0.60241	0.45	0.819672	33	50	83	(2, 3, 0, 4, 1)	10
0.601093	0.209677	0.601093	-0.21667	1.803279	73	110	183	(2, 0, 3, 1, 4)	11
0.591549	0.62	0.591549	0.516667	0.688525	29	42	71	(1, 3, 0, 4, 2)	12
0.589286	0.345912	0.589286	-0.91667	2.704918	115	165	280	(3, 2, 4, 0, 1)	13
0.58701	0.406353	0.58701	-10.2333	15.70492	674	958	1632	(2, 3, 4, 0, 1)	14
0.58626	0.409717	0.58626	-16.8667	24.90164	1072	1519	2591	(1, 3, 0, 2, 4)	15

2.2 Preprocessing – Approach 2

In the second approach of the project, we focused on building a modular and fully automated preprocessing pipeline to prepare the EEG recordings for pattern analysis. The process began with converting raw EEG files from .mat format to .npy, enabling efficient and consistent data loading using Python. Each subject was assigned to either the ADHD or CONTROL group, and their files were saved in a structured directory.

We then applied several signal cleaning steps. First, we used a bandpass filter (typically 1–45 Hz) to remove low-frequency drift and high-frequency noise outside the standard EEG range. Next, we performed Independent Component Analysis (ICA) to detect and eliminate common artifacts such as eye blinks and muscle activity. Specifically, we applied the FastICA algorithm and measured the kurtosis of each component—assuming components with high kurtosis (above a predefined threshold) represent noise-like artifacts. These components were automatically removed by zeroing them out before reconstructing the cleaned signal. This method allowed for unsupervised and effective noise removal.

After artifact removal, we normalized each EEG channel to have zero mean and unit variance to ensure consistency across subjects.

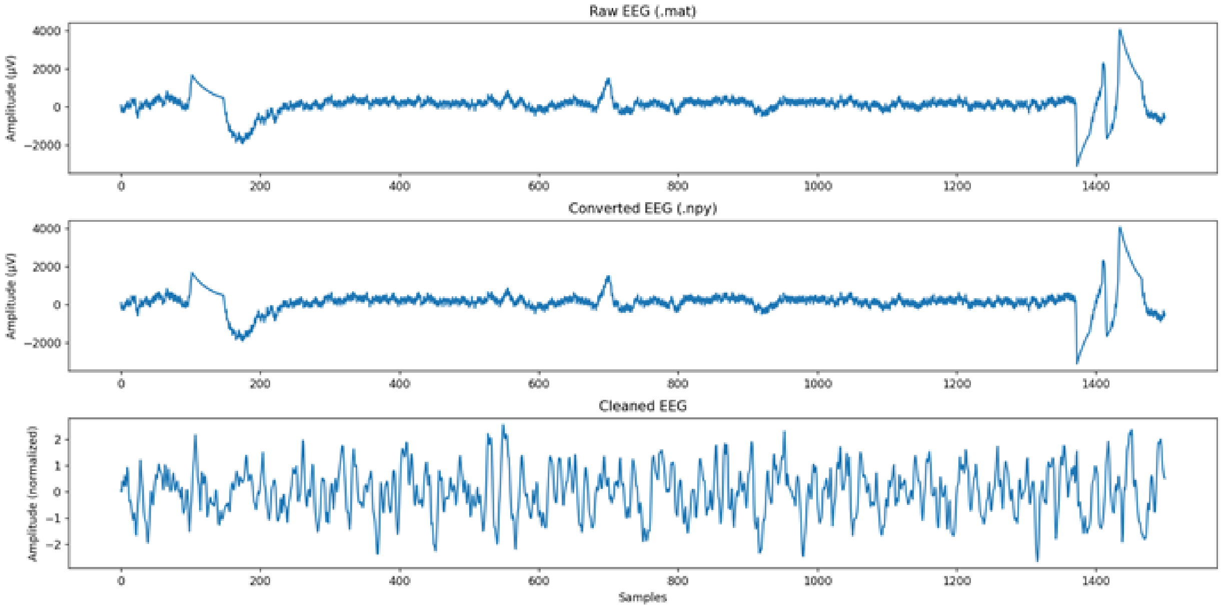
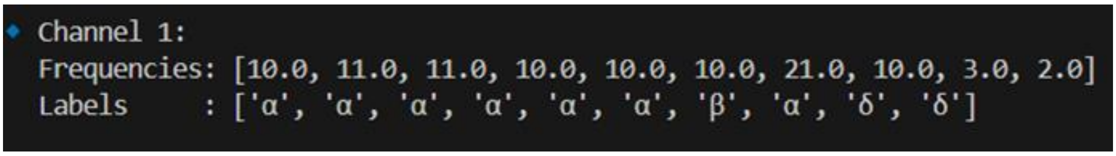


Fig. 1: Example EEG signal before and after conversion, cleaning and normalizing, showing the effect of preprocessing

Following the cleaning stage, we segmented each EEG recording into overlapping windows, using various durations and overlap settings - such as 1-second windows with 50% overlap - as a representative example. These short, overlapping windows allowed us to capture localized EEG activity over time, while maintaining sufficient continuity for pattern recognition.

For each window and channel, we computed the dominant frequency using Welch's method and mapped these numeric frequencies to symbolic brainwave bands (δ, θ, α, β, γ). This dual representation—numerical for order analysis and symbolic for band labeling—was specifically designed to support the next phase, which uses the Order Preserving Matching (OPM) algorithm as the core of our project.



```
🔹 Channel 1:
  Frequencies: [10.0, 11.0, 11.0, 10.0, 10.0, 10.0, 21.0, 10.0, 3.0, 2.0]
  Labels     : ['α', 'α', 'α', 'α', 'α', 'α', 'β', 'α', 'δ', 'δ']
```

Fig. 2: *Example of the first 10 frequency values in a single EEG channel. The top row shows the dominant numerical frequencies per time window, while the bottom row presents the corresponding symbolic brainwave labels (δ, θ, α, β, γ) used in pattern detection.*

By the end of this preprocessing stage, we had a clean, structured, and fully labeled dataset ready for OPM-based pattern detection and group comparison.

3. Challenges and Solutions During the Project

Throughout the project, we encountered several technical and methodological challenges that required creative solutions and adjustments to our workflow:

3.1 Approach 1

Integrating the AMICA Algorithm into a Modern Environment

One of the main challenges in this approach was integrating the original AMICA executable into a modern Python-based pipeline. Since AMICA is not available as a standard Python library, we had to compile it externally and build a stable wrapper that allowed it to be invoked from within the overall code. In addition, we created a dedicated dependencies file along with environment variable settings, ensuring consistent execution across machines, even when the AMICA binary required specific runtime conditions. This setup enabled portability, reproducibility, and ease of deployment

Improving Runtime Performance

Another significant challenge was the long runtime required for comparing patterns. When attempting to match motifs using naive brute-force search, the processing time increased dramatically with dataset size. To address this, we implemented a hash table mechanism that stores each detected OPM pattern, avoiding redundant comparisons. This optimization reduced the runtime from days to minutes and enabled efficient motif scanning even across large datasets

File Management and Memory Optimization

Running AMICA and motif searches generated large output files that became difficult to manage. To handle this, we created lightweight result files that include only the necessary outputs for GUI display and analysis, excluding unused data. This significantly reduced disk usage while preserving analytical capabilities

Supporting Diverse Signal Processing Approaches

An additional complexity stemmed from the realization that relying on a single motif representation was insufficient. For example, matching raw band-pass signals might miss patterns that are better captured by the Hilbert envelope or signal trends. To overcome this, we extended the system to support eight distinct matching configurations, combining Independent Components (ICs) and raw Channels with four different processing modes — Raw, Raw+Trend, Hilbert, and Hilbert+Trend. This required a flexible code design but allowed more comprehensive and meaningful pattern discovery

Pipeline Division into Two Stages

Due to the heavy computational cost of AMICA, we decided to divide the pipeline into two clear stages. The first stage performs the ICA decomposition and stores the results for reuse, while the second stage focuses on motif matching once the ICA outputs are validated. This separation allowed for more efficient execution, easier debugging, and modular experimentation, including use cases where only one of the stages needs to be rerun

Efficient Memory Management and Write-Back Strategy for Large-Scale Pattern Matching in EEG Data

During the development of the pattern identification system, we encountered a significant engineering challenge related to memory management. The difficulty arose from the increasing volume of data required for processing and the complexity of pattern matching operations across large sets of EEG recordings.

Initially, the system was designed to store all intermediate data structures—such as pattern tables, subject identifiers, time indices, and symbolic OPM sequences—in memory (RAM) to enable fast access and simplified management. However, as the dataset grew, especially under repeated comparisons of patterns across multiple signals, memory usage became excessive and, in many cases, unstable. This resulted in system slowdowns, crashes, and an inability to scale to larger datasets.

The challenge became even more acute when the system was required to repeatedly update the same data entry—such as when a specific pattern occurrence was revisited or recalculated. Without a consistent mechanism for detecting and updating existing entries, the system risked introducing duplicates or overwriting data unintentionally.

To address this issue, we implemented a **write-back memory management strategy**, inspired by caching mechanisms commonly used in Intel CPU architectures. In this strategy, write operations to storage are deferred and managed through an internal buffer (or cache), and only written to permanent storage when certain conditions are met (e.g., eviction, synchronization, or completion of processing). Each entry in memory was tracked using two control bits: A **valid bit**, indicating whether the stored value is current and usable, and a **dirty bit**, indicating whether the value has been modified since it was last written to disk.

This approach allowed the system to determine whether a previously written row needed to be updated (i.e., overwritten) or left unchanged. For example, when a pattern occurrence was revisited, the system would check whether the record was already marked as valid and whether it had changed (dirty = 1). Only then would an intentional overwrite operation be triggered, ensuring correctness while avoiding unnecessary I/O.

In parallel, we adopted a lightweight file architecture that separated large-scale data (e.g., full pattern appearances) from compact statistical summaries. For instance, files such as `pattern_metrics.parquet` contained only essential statistical descriptors, allowing the system to remain efficient while minimizing memory footprint.

By combining this rule-based memory management scheme with an efficient write-back strategy and file structure separation, we were able to overcome the limitations of in-memory processing. This solution ensured stable performance, reduced memory usage, and preserved data integrity even under intensive pattern matching operations across large EEG datasets.

3.2 Approach 2

Data Format Conversion

The original EEG recordings were provided in .mat files (used by MATLAB), which are not easily processed in Python. We implemented a preprocessing step to convert them to .npy format, while validating that each file contained usable data. Some files were skipped due to missing or irrelevant content.

Choosing a Matching Threshold

When comparing patterns between subjects, we had to define a similarity threshold—how close two sequences must be to be considered a match. A low threshold resulted in too many false positives, while a high threshold filtered out meaningful patterns. We tested several threshold values (e.g., 0.8, 0.9, 1.0) to find a good balance between sensitivity and specificity.

Label Indexing Problem

At one point, we tried to convert frequency labels (e.g., α, β, α) into numerical values (e.g., 0,1,0) to enable easier comparisons. However, this caused issues with the OPM method, which is based solely on relative order. For instance, both sequences (0,1,2) and (1,2,3) would produce the same OPM structure, even though they represent different label patterns.

Solution: Dual Representation

To solve this, we separated the representation into two parallel arrays: one containing the actual frequency values, and the other containing the symbolic band labels (e.g., α, β). This allowed us to:

Use the frequency array to compute the OPM structure.
Use the label array to ensure exact label sequence matching.

This approach preserved the advantages of OPM while ensuring that only true label-pattern matches were accepted.
Addressing these challenges helped us build a more reliable and accurate pattern comparison framework and highlighted the importance of combining advanced algorithms with thoughtful data representation.

4. Solution

4.1 Solution – Approach 1

In the analytical phase of Approach 1, our goal was to examine whether the waveform structure itself—not just the frequency sequence—could constitute a distinguishing pattern between children with ADHD and those in the control group. To that end, we focused on theta-band activity and explored whether recurring oscillatory structures appeared predominantly in the ADHD group

The detection process relied on a combination of time-domain and phase-domain analysis. After extracting Independent Components using AMICA, we applied a band-pass filter in the 4–8 Hz range to isolate theta activity. We then used the Hilbert transform to compute the signal's amplitude envelope and instantaneous phase. Bursts were identified using strict criteria: envelope values exceeding 1.5 standard deviations above the mean, and phase progression of at least 6π—ensuring the inclusion of at least three complete theta cycles

From each burst, we extracted a single pattern consisting of three cycles located between two consecutive phase transitions. This pattern was encoded as an OPM structure that captures the relative rank order of signal values across time. This representation preserved the shape of the waveform regardless of amplitude variation. Each pattern was stored in a hash table to avoid redundant comparisons, providing high computational efficiency

Each pattern was evaluated across eight distinct configurations, combining two processing axes: the source (Independent Component vs. raw channel) and the transformation type (Raw, Raw+Trend, Hilbert, Hilbert+Trend). Our objective was to find recurring waveform patterns that appeared consistently in one group and were rare in the other

Every detected pattern underwent a two-phase validation: first, we verified its prevalence within the source group (ADHD or control), and then we assessed its presence in the opposite group.

Through this integration of precise cycle detection, strict filtering, and structure-preserving hashing, we succeeded in identifying motifs that appeared with much higher frequency in one group only—supporting the hypothesis that waveform periodicity itself may serve as a neurophysiological signature distinguishing children with ADHD from the control population

4.2 Solution – Approach 2

Following the preprocessing phase, we entered the core analytical stage of our project: the detection and comparison of recurring EEG patterns across the ADHD and Control groups. At the heart of this process lies the Order Preserving Matching (OPM) algorithm, which was central to our methodology.

The goal was to identify frequency band patterns—such as sequences like ['α', 'β', 'β']—that appear consistently across individuals within one group, while being rare or absent in the opposite group. To achieve this, we applied a dual filtering strategy:
First, we represent each short sequence by its OPM structure, which encodes the internal order of frequency values.
Then, we combined this with the exact symbolic labels (δ, θ, α, β, γ), ensuring that both the order and the type of brainwave bands matched. This joint criterion increased the precision of our pattern matching, allowing us to identify recurring structures that are both numerically ordered and semantically meaningful.

To ensure robustness, we applied this analysis bidirectionally:

- First, we searched for patterns that were common in the ADHD group and tested whether they also appeared in the Control group.
- Then, we reversed the process by identifying frequent patterns in the Control group and evaluating their presence in the ADHD group.

The pattern discovery process began by scanning each subject's EEG recordings—per channel and per time window—and extracting short, overlapping sequences of frequency labels.
We tested multiple pattern lengths (typically 3–6), and we also applied this analysis across different window durations—such as 1-second and 2-second windows—to capture both fast-changing and slower-developing EEG dynamics. Each candidate sequence was transformed into its OPM representation and labeled according to its brainwave band symbols.

Patterns were then aggregated across all subjects in the same group. A pattern was considered group-representative if:

- It appeared in at least 80% of the subjects in the group,
- And it exactly matched both the symbolic sequence and the OPM order,
- And it occurred in the same EEG channel across subjects.

Once prominent patterns were identified, we moved to the cross-group comparison step. Each pattern was searched in the opposite group, and a match was accepted only if the symbolic sequence and the OPM structure had at least 80% similarity.

If a pattern was also common in the opposite group, it was excluded. Instead, we focused on the most distinctive patterns—those with the largest gap in appearance rates between the two groups. For example, a pattern that appeared in 85% of ADHD subjects but only 50% of Control subjects would be considered group-specific.

By executing this process in both directions and applying strict dual-matching criteria, we ensured that the final set of patterns was both robust and indicative of group-level differences. This analytical design highlights the power of OPM as a novel approach to symbolic EEG pattern analysis and group classification.

5. Pattern Detection Results

Approach 1

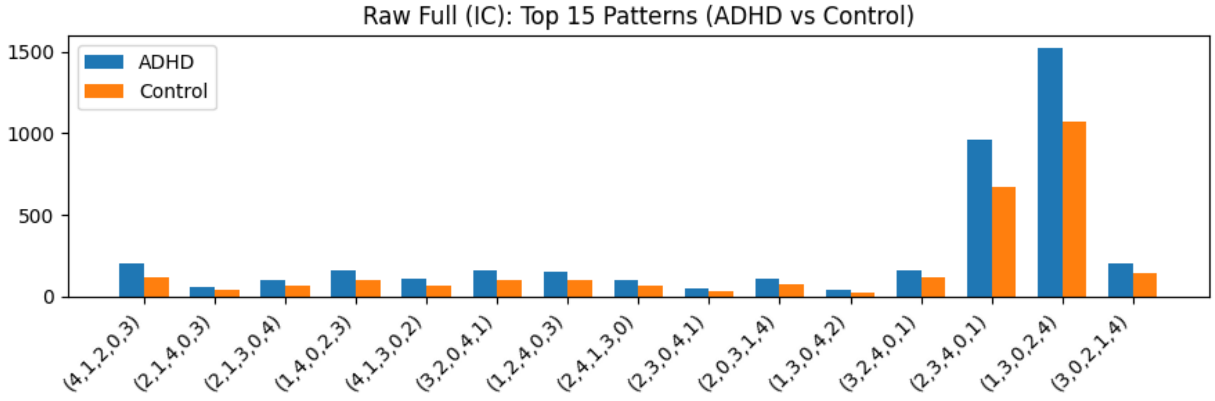
After completing the OPM-based analysis on individual EEG waveform structures, we compiled the most frequent and distinctive symbolic patterns derived from single-cycle morphologies across subjects. For each tested configuration—defined by the signal source (independent component or raw channel) and the waveform representation type (Raw, Trend, Hilbert, Hilbert + Trend)—we identified patterns that met the detection criteria within the ADHD group and evaluated their recurrence in the control group, and vice versa.

Unlike approaches that rely on predefined time windows, this method focuses on phase-defined waveforms extracted based on strict amplitude and phase progression criteria. Each selected waveform was transformed into an OPM sequence, and matches were identified based solely on structural similarity using a rule-based comparison mechanism.

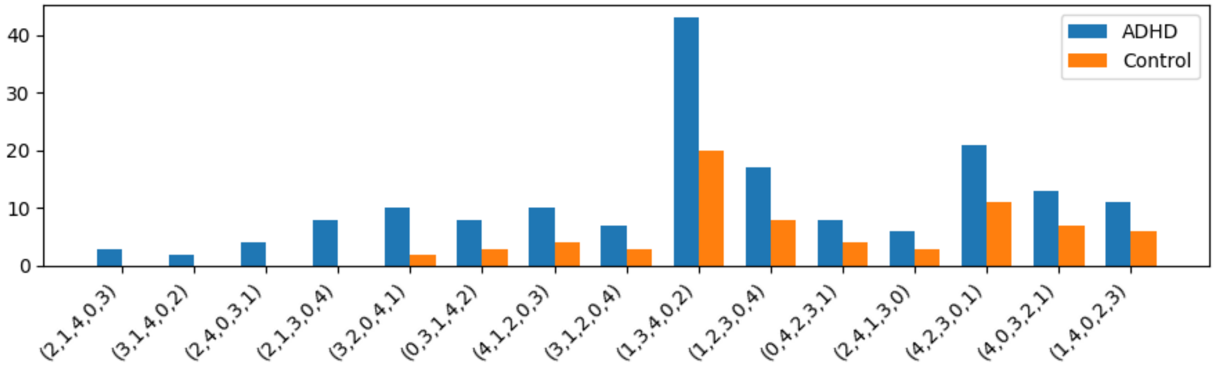
We visualized the top-ranked patterns based on their relative frequency gap between groups, emphasizing those that appeared in a large proportion of subjects in one group (≥90%) but were absent or rare in the opposite group. This direct and deterministic comparison allowed us to isolate patterns that consistently differentiated between ADHD and control participants, without applying statistical thresholds or probability models.

To ensure robustness, the analysis was repeated across all eight processing configurations. The resulting set of patterns provides insight into structural motifs that recur more frequently in one group, and demonstrates the potential of morphology-based analysis as a complementary diagnostic tool.

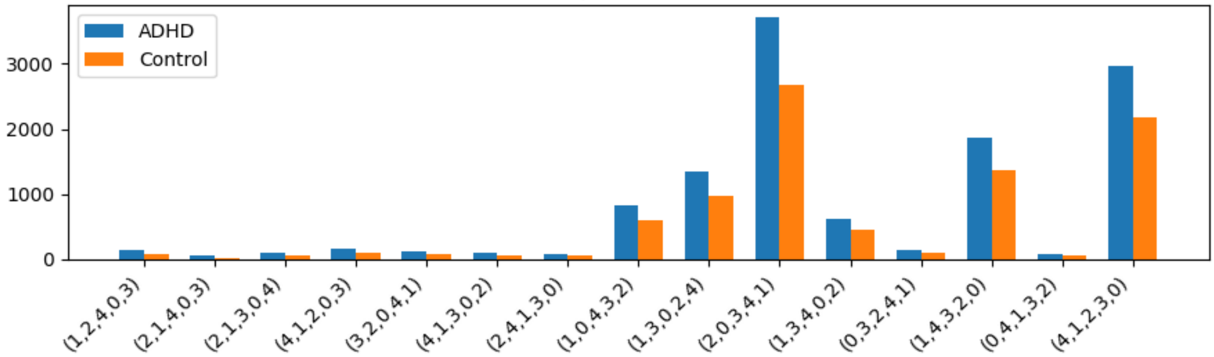
Here are the graphs of the best patterns across the PPV index (the primary metric among global health organizations) for the 8 patterns of approach 1:



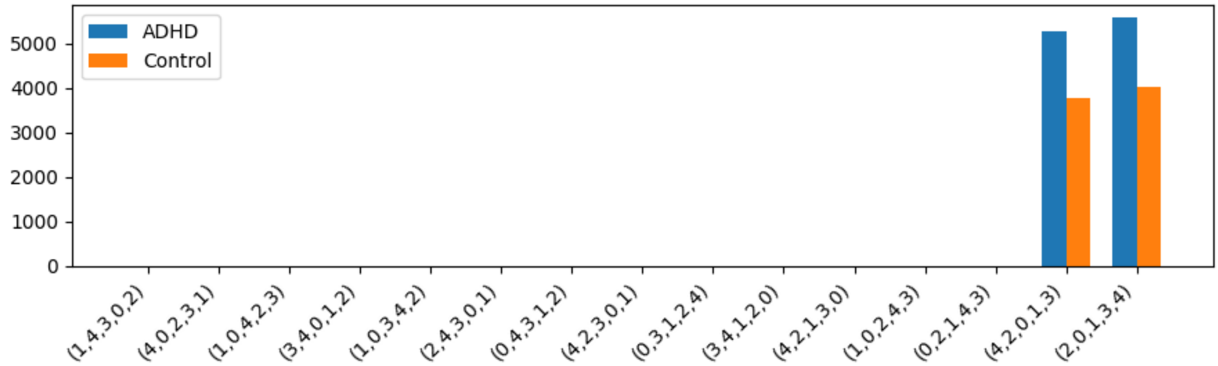
Raw Trend (IC): Top 15 Patterns (ADHD vs Control)



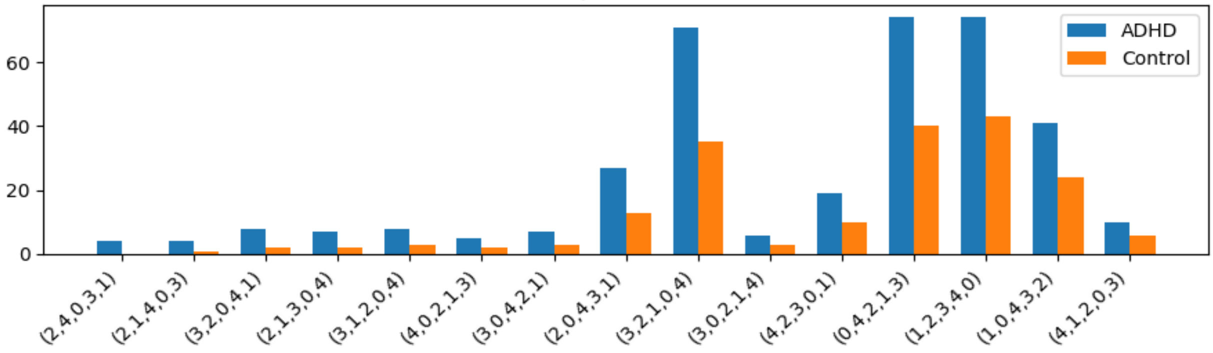
Raw Full (Channel): Top 15 Patterns (ADHD vs Control)



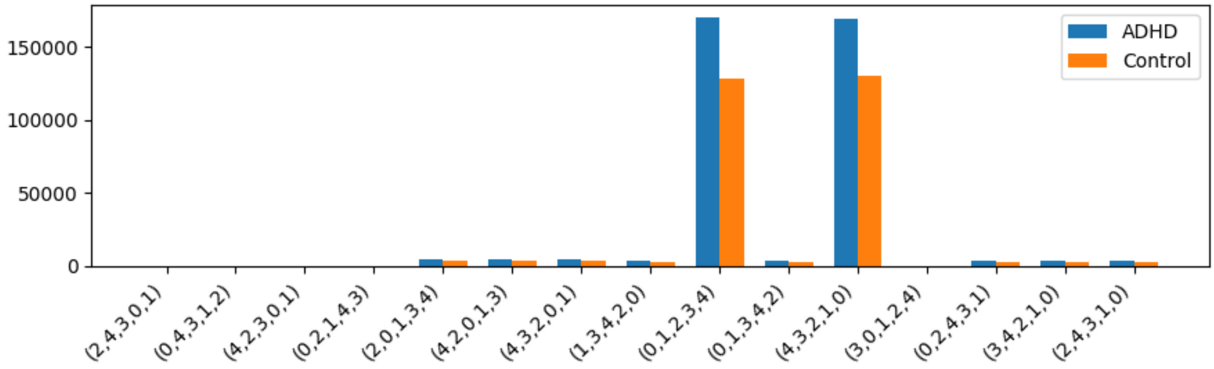
Hilbert Full (IC): Top 15 Patterns (ADHD vs Control)

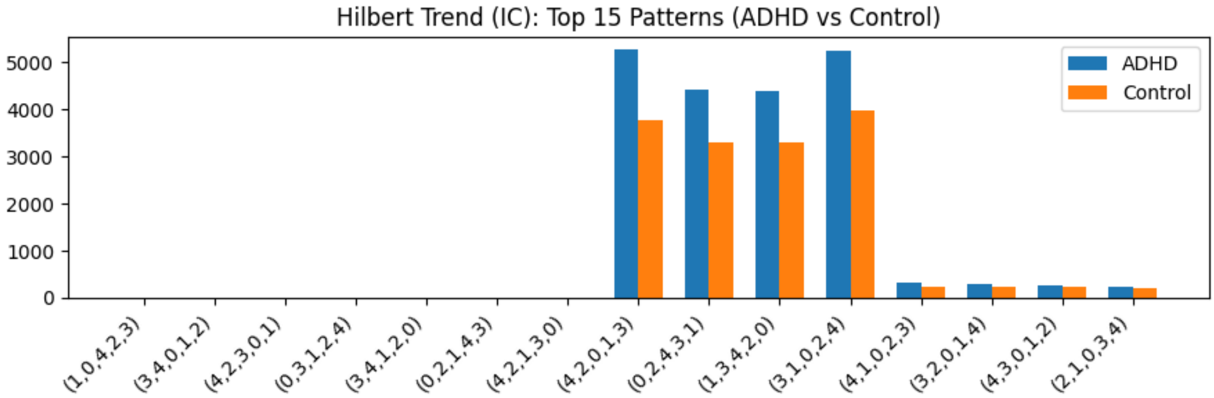


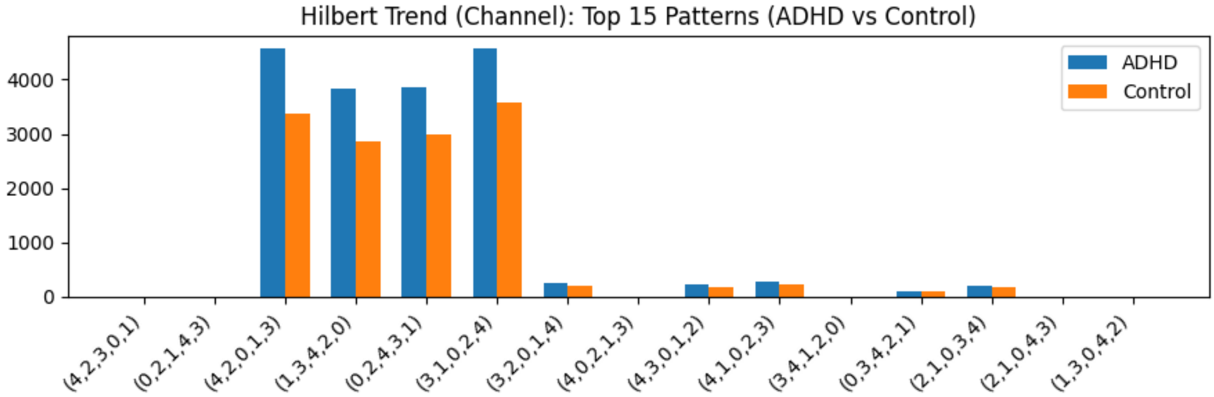
Raw Trend (Channel): Top 15 Patterns (ADHD vs Control)



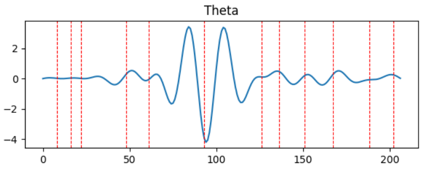
Hilbert Full (Channel): Top 15 Patterns (ADHD vs Control)

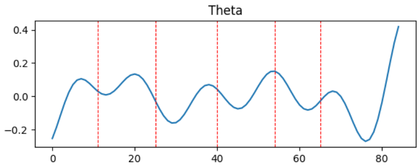


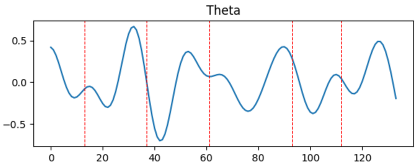




Wave structure found for illustration:







After completing the OPM-based pattern analysis, we compiled the most representative and distinctive EEG patterns for each direction of comparison. For each tested configuration (e.g., pattern length, time window size), we recorded the patterns that passed all filtering criteria within the source group and then evaluated their match rates in the opposite group.

We visualized the top patterns based on their appearance gap between groups—highlighting

those that were highly prevalent in one group (≥80% of subjects in the same EEG channel) and significantly less common in the other. This gap serves as an indicator of group specificity and strengthens the validity of the identified pattern as a potential biomarker.

To provide a comprehensive view, we also repeated this analysis using different time window lengths (e.g., 1s and 2s), This allowed us to uncover additional patterns that may only emerge at different temporal scales, and to increase the chances of capturing group-specific dynamics that vary in duration.

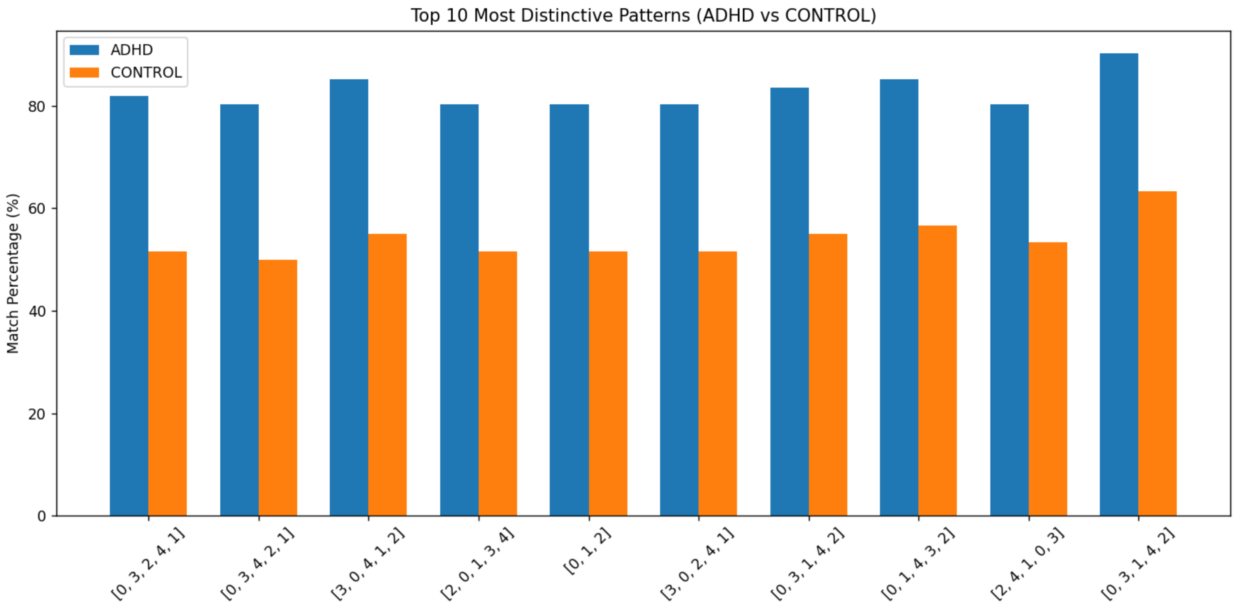


Fig. 3: Top 10 most distinctive symbolic EEG patterns found in the ADHD group (using 1s windows, 50% overlap), and their appearance rate in the Control group.

All top 10 patterns exhibited strong presence in the ADHD group (typically over 80%) and substantially lower presence in the Control group (mostly around 50%–55%).

Although we also tested longer patterns (e.g., length 6) during the OPM analysis, none of them ranked among the top 10 most distinctive. This may indicate that longer symbolic sequences are either less robust or less consistently shared across ADHD subjects when using 1-second windows.

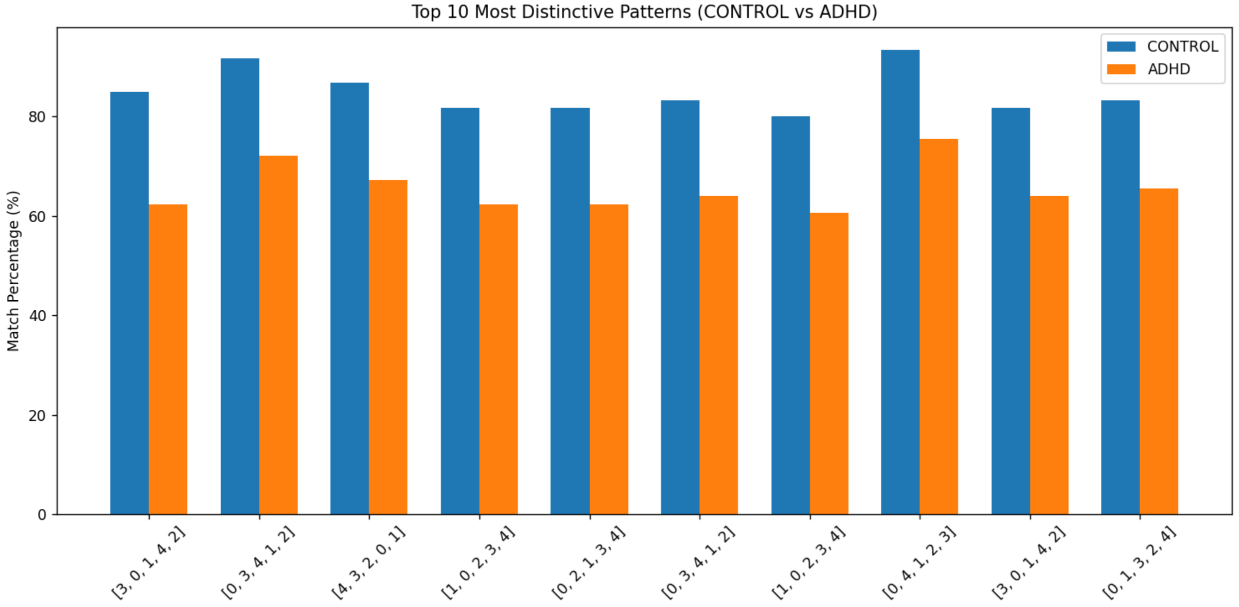


Fig. 4: Top 10 Most Distinctive Patterns (CONTROL vs ADHD) – 1s window

All top 10 patterns in this direction were highly frequent across Control subjects (typically above 80%) and significantly less common in the ADHD group (mostly between 60%–70%). This highlights distinct EEG dynamics in the Control group that are less consistently observed in ADHD recordings, using a 1-second time window with 50% overlap.

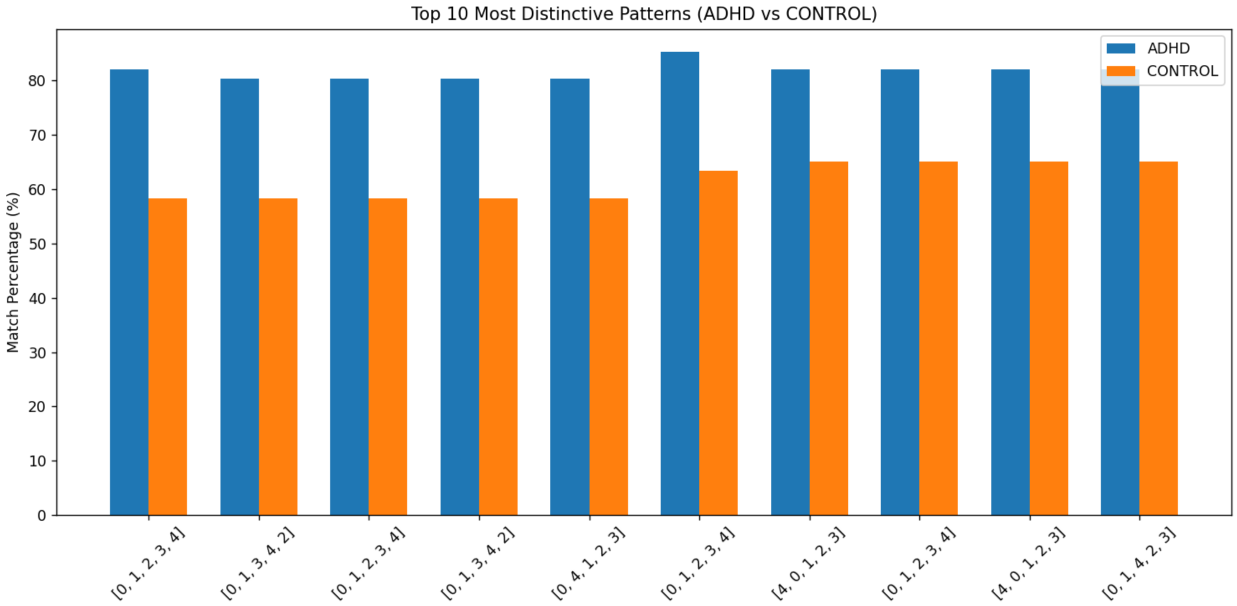


Fig. 5: Top 10 Most Distinctive Patterns (ADHD vs CONTROL) – 2s window

All top 10 patterns shown here were frequently observed in the ADHD group (typically above 80%) and less common in the Control group (mostly around 58%–65%).

Compared to the 1-second window configuration, this setting yielded patterns with smaller gaps between groups, suggesting that longer time windows may reduce the distinctiveness of ADHD-related EEG sequences—possibly due to increased variability over extended intervals.

To better illustrate these differences, we simulated the waveform structures of the six most distinctive patterns (Δ ≥ ~30%) using synthetic EEG-like signals:

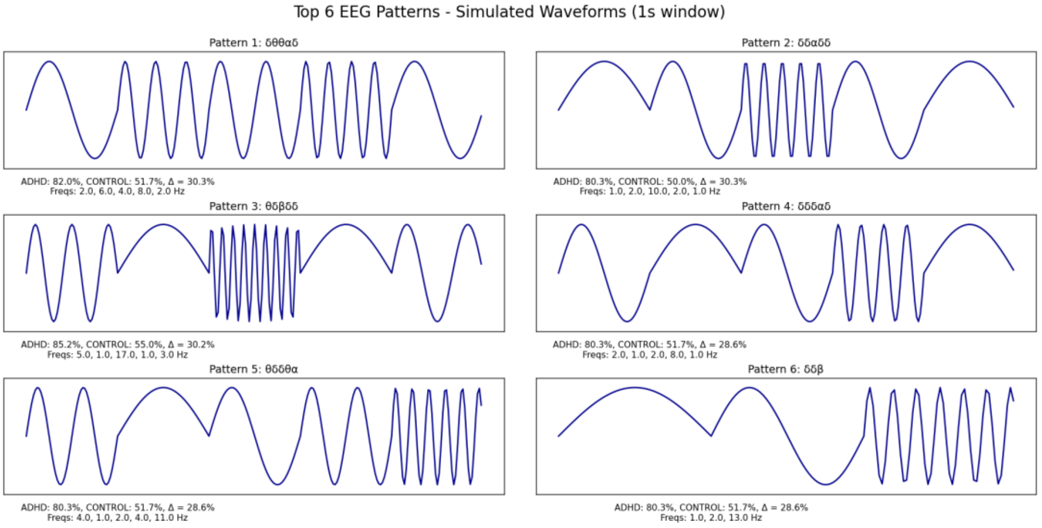


Fig. 6: Top 6 EEG Patterns – Simulated Waveforms (1s window)

This figure shows simulated versions of the six most distinctive EEG patterns (based on the difference between ADHD and Control groups, Δ) identified using a 1-second window.

Each waveform was generated by summing sine waves at the listed frequencies, providing a visual approximation of how these patterns might appear in real EEG data and highlighting the range of brainwave activity captured by the OPM method.

6. Conclusions and Achievement of Research Goals

The second phase of our project fulfilled the core research goal set in the first phase: identifying recurring EEG patterns that are characteristic of either the ADHD or the Control group. Through a modular preprocessing pipeline and systematic use of the OPM algorithm, we uncovered symbolic frequency patterns that recur across subjects in one group while appearing less frequently in the other. Although some patterns—like ('δ', 'δ', 'δ', 'δ', 'δ')—appeared in both groups, their prevalence was found to be channel-specific, with notable differences such as higher frequency in F3 (associated with ADHD) for ADHD subjects, and in other channels for the Control group. This highlights the importance of spatial localization in EEG analysis.

While the strict 90% IOU threshold defined in Phase 1 was not consistently met, we successfully identified multiple patterns with large intergroup appearance gaps (20–30%), particularly in frontal and central EEG channels. These findings support the hypothesis that distinct, interpretable symbolic patterns exist and are distributed differently across ADHD and non-ADHD brains. A supplementary file listing all identified patterns, channels, and group match statistics has been added to the project folder for future reference and exploration.

7. Personal Reflections & Future Work

During this phase of the project, we encountered unexpected challenges that prompted us to refine our methodology and deepen our understanding of the dynamics of EEG signals. One key insight was the importance of combining the OPM structure with symbolic frequency labels to improve the specificity of the patterns. We found that in Approach 1, which examines the waveform structure itself, there were patterns that differed significantly between the ADHD group and the control group. We also found that one-second time windows yielded particularly unique results. These lessons will guide the next phases of the research, especially in assessing clinical utility and expanding the framework for pattern recognition.

After answering the research question, we can think about future research: Since in Approach 1 a significant gap was found in some of the patterns between the ADHD group and the control group, in the event that we take these patterns and create a sequence of waves from them and search, as in Approach 2, for the sequence with the highest frequency among ADHD subjects, could we use this to find a sequence of waves that is a significant marker for the presence of ADHD? Do such sequences exist? And what if we take waves from IC that we found in Approach 1 and search for waves that appeared after them in other ICs from other channels? Would we find brain communication activity in this way?