Software Engineering Department
Braude College

Capstone Project Phase B

EEG Classification Using Text Compression

25-1-R-2

User Guide

Maintenance Guide

By:- Mohammad khateeb
Jad taha

Advisors:- Dr. Samah Idrees Ghazawi
Dr. Anat Dahan

Link to github:-
https://github.com/mhmdkh1905/EEG-recordings.git

Contents

1. Introduction

Electroencephalography (EEG) signals provide a rich source of information for understanding brain activity, especially in distinguishing between clinical and non-clinical populations such as individuals with Attention Deficit Hyperactivity Disorder (ADHD). Traditional classification methods often rely on supervised machine learning models, which require significant labeled data, feature engineering, and domain expertise. However, EEG signals are inherently complex, noisy, and high-dimensional, making these methods both resource-intensive and potentially limited in generalizability.

In this project, we propose an alternative approach based on text compression techniques, specifically the use of the Normalized Compression Distance (NCD) computed via the ZB2 compression algorithm. Our hypothesis is that EEG signals, when transformed into text-like representations, can be meaningfully compared using compression-based similarity measures. This approach avoids the need for supervised learning and instead focuses on structural similarities between signals, leveraging the theory that more similar signals will compress better together than dissimilar ones.

One of the key motivations for using compression-based analysis is its ability to expose recurring patterns and structural features embedded in the data. Compression algorithms inherently identify redundancy and repeated sequences in signals as part of their encoding logic. In the context of EEG analysis, this characteristic allows us to capture hidden regularities that may differ between ADHD and control groups. By measuring how efficiently different EEG sequences compress relative to one another, we can infer underlying signal complexity and structural similarity — without requiring feature extraction or domain-specific assumptions.

The primary goal of our project is to classify EEG signals from children diagnosed with ADHD and those from typically developing controls by analyzing their compressed representations. By relying on NCD and statistical similarity, we aim to develop a scalable, unsupervised methodology that can offer insights into brain signal patterns with minimal preprocessing and no model training.

1.1 Scope of This Document

This document outlines the second phase of our capstone project, focusing on the implementation and evaluation of our proposed solution for EEG classification using text compression. It builds upon the conceptual design and objectives detailed in Phase A and presents the technical processes, data preparation strategies, evaluation metrics, and experimental outcomes of our approach.

We describe the dataset used, the preprocessing steps applied to the EEG recordings, the methodology for dividing and comparing signals, and the results obtained from our classification pipeline. This report is intended for researchers and practitioners in the fields of biomedical signal processing, data compression, and computational neuroscience. It provides a comprehensive overview of the practical

work carried out, along with the challenges encountered and the lessons learned throughout the development and experimentation stages.

2. General Description

The goal of our project is to classify EEG signals recorded from children diagnosed with Attention Deficit Hyperactivity Disorder (ADHD) and from typically developing controls by analyzing the structural characteristics of their brainwave patterns. The system relies on the idea that these signals, once transformed into textual representations, can be meaningfully compared using a compression-based similarity measure known as the Normalized Compression Distance (NCD). By dividing each EEG signal into smaller parts and calculating the NCD between parts from different participants, we are able to assess the degree of similarity between signals. We then use these similarity values to determine the most likely classification of each signal based on its proximity to other known examples.

The system is designed to be simple, interpretable, and adaptable. It includes several core components such as signal filtering, brainwave frequency region extraction, segmentation of signals into fixed-length parts, and comparison of these parts using a dedicated compression algorithm called ZB2. The intended users of this method are researchers and professionals involved in the analysis of EEG signals, particularly in clinical and cognitive studies. The modular nature of our implementation allows for flexibility in adapting the system to various datasets, regions, and evaluation methods, making it a useful tool for examining patterns of brain activity in a structured and reproducible way.

3. Proposed Solution

The proposed solution is based on analyzing and comparing EEG brainwave signals using a compression-based similarity approach. Instead of relying on mathematical models or manual feature extraction, we focus on measuring the structural similarity between signals by compressing them and evaluating how efficiently they can be represented together. This is achieved using the Normalized Compression Distance (NCD), a metric that quantifies the similarity between two sequences based on their compressed sizes. The underlying idea is that signals with similar patterns will compress better together, resulting in a lower NCD value.

To implement this approach, we first preprocess the EEG recordings to reduce noise and isolate relevant brainwave frequency bands. Each signal is filtered and then segmented according to brainwave regions such as Alpha, Beta, Gamma, and others. These segments are further divided into fixed-length parts of 1000 characters, representing sections of the signal. For classification, we compare each unknown signal with a set of known signals from both the ADHD group and the control group. The NCD is computed for each pair of parts using a dedicated compression algorithm (ZB2), and the results are summarized using statistical measures. Based on these similarity scores, we determine whether the signal is more likely to belong to the ADHD group or the control group.

This solution allows us to classify EEG signals without relying on labeled training data or machine learning models. It is particularly well-suited to scenarios where interpretability, reproducibility, and scalability are important. The modular nature of the process enables researchers to adapt the workflow to different datasets, brain regions, or compression strategies, making it a flexible framework for signal analysis.

3.1. Dataset

The dataset used in this project consists of EEG recordings collected from a total of 121 children, including 61 diagnosed with Attention Deficit Hyperactivity Disorder (ADHD) and 60 typically developing children serving as the control group. All participants were between the ages of 7 and 12 and included both boys and girls. The ADHD group was diagnosed by a certified psychiatrist according to DSM-IV criteria and had been treated with Ritalin for up to six months. The control group was selected to ensure no history of psychiatric or neurological disorders, including epilepsy or high-risk behaviors.

EEG signals were recorded using the international 10–20 system across 19 standard channels: Fz, Cz, Pz, C3, C4, T3, T4, Fp1, Fp2, F3, F4, F7, F8, P3, P4, T5, T6, O1, and O2. The sampling frequency for all recordings was 128 Hz. The reference electrodes were A1 and A2, placed on the earlobes, in accordance with standard EEG recording procedures.

During the EEG recording session, participants engaged in a continuous visual attention task. The task involved observing cartoon characters and counting them within each presented image. The number of characters varied randomly between five and sixteen. As soon as the child responded, a new image appeared immediately, ensuring a continuous and uninterrupted stream of stimuli throughout the recording session. This setup was designed to maintain consistent attention and maximize the quality of the brainwave data collected during task performance.

The dataset is stored in CSV format, with each file representing the full EEG recording of a single participant. Each column corresponds to one of the 19 EEG channels, and each row represents a single time sample. These files serve as the input for the preprocessing and analysis stages in our proposed solution.

3.2. Methodology and Algorithmic Structure

Our solution is based on analyzing the structural content of EEG signals by comparing them using the Normalized Compression Distance (NCD). To enable this comparison, we first process each EEG signal and convert it into a textual representation. Each signal is filtered to remove noise, focusing on the 1–40 Hz frequency range, and then segmented into distinct brainwave regions including Delta, Theta, Alpha, Beta, and Gamma. These regions are generated using time-based sliding windows of various sizes, such as 5, 6, 8, and 10 seconds, allowing us to capture different levels of temporal resolution.

Once the brainwave-specific segments are prepared, we divide each resulting text sequence into equal parts of 1000 characters. The next step involves computing the NCD between all parts of one participant and all parts of another participant using the ZB2 compression algorithm. This produces a set of NCD values for each participant pair. For every pairwise comparison, we calculate the minimum, average, and median NCD values, which summarize the similarity across all parts. The median value is selected as the main indicator for the final classification decision, as it tends to offer a robust measure that is less affected by outliers.

To perform classification, we focus on one participant at a time. We take that participant's set of NCD similarity scores with all other participants and compute the median of those values. This median represents the participant's overall structural similarity threshold. Then, for each of the other participants, we compare their NCD score to this threshold: if the score is less than or equal to the median, the other participant is considered to belong to the same group; otherwise, they are considered to belong to the opposite group. This logic is applied iteratively to all participants to determine predicted group membership based on pairwise similarity.

The process is repeated across all 19 EEG channels for all participants, and the results are organized in structured folders corresponding to each version and region. The comparison results for each channel and each version are saved for further analysis. This systematic methodology enables us to determine whether a signal is more similar to the ADHD group or the control group, thereby supporting classification without relying on predefined models or training procedures. A simplified flowchart of the method is presented in Figure 1, outlining the key steps from preprocessing to classification output.

A detailed overview of the complete EEG classification pipeline is presented in Figure 2. The diagram captures the key steps starting from filtering and segmentation, through region extraction and part-wise NCD calculation, and concluding with statistical summarization and the final classification logic. This visual guide clarifies how minimum, average, and median NCD values are computed for each participant pair and how these summaries are used to determine group assignment based on structural similarity.

Figure 1. EEG signal processing and classification pipeline using NCD and ZB2 compression.

The figure illustrates the sequential steps applied in the project, starting from data collection and ending with classification and evaluation of ADHD versus control EEG signals.

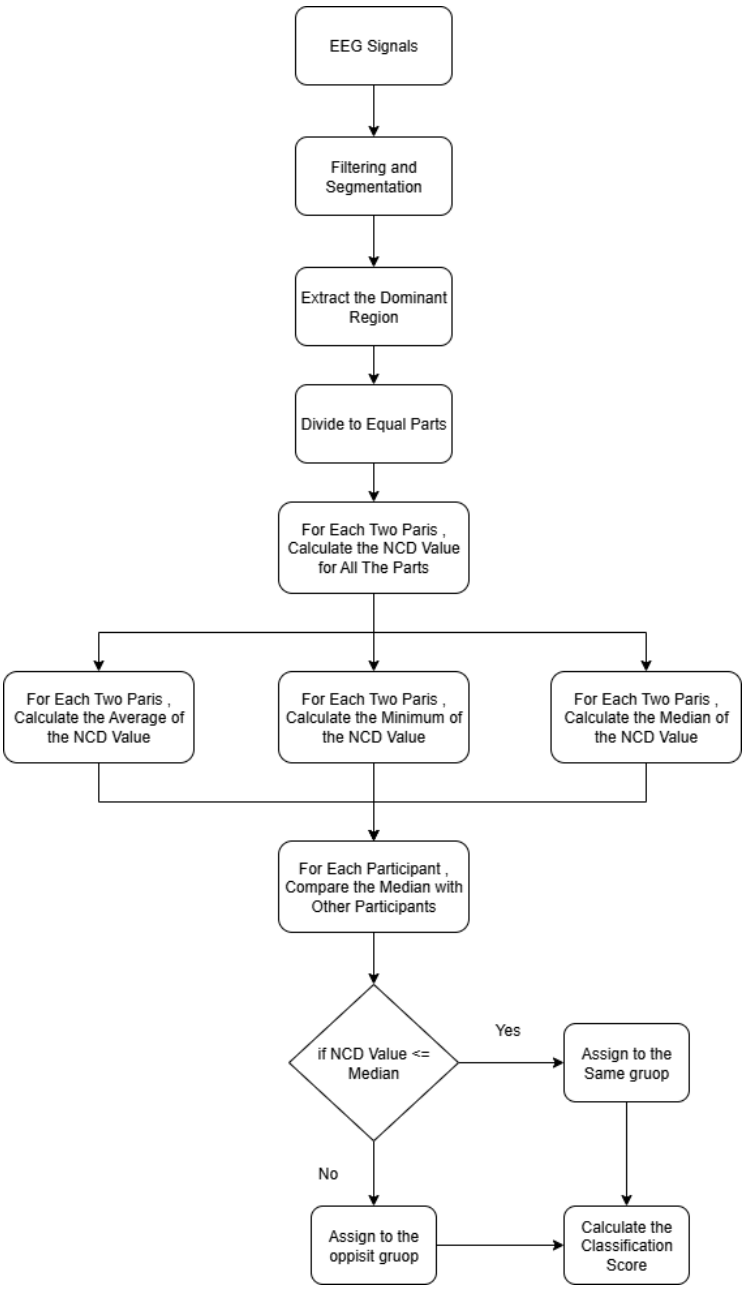

Figure 2. Flowchart summarizing the full EEG classification process using NCD, from preprocessing to final group assignment.

3.3. Architecture Description

The architecture of our system is organized into modular processing stages that reflect the sequential structure of the EEG classification workflow. All components were implemented using Python in Google Colab, with data stored and accessed directly from Google Drive. This setup allowed us to manage a large number of files efficiently, while maintaining a clear separation between each processing phase.

At a high level, the system is composed of the following layers: data acquisition and preprocessing, frequency region extraction, segmentation and formatting, compression-based comparison, and result aggregation. Each step was implemented in

separate scripts to enable reusability and simplify debugging. The filtered EEG files are stored in dedicated folders according to the participant group (ADHD or control). These files are then passed to the region extraction module, which generates new versions for each desired brainwave region and window size.

The segmentation layer splits each brainwave region into parts of fixed length (1000 characters), and stores them in structured subdirectories labeled by channel and version. The comparison layer iteratively selects a signal from one participant and computes the Normalized Compression Distance (NCD) between all of its parts and all parts of another participant using the ZB2 compression algorithm. The results of each comparison are saved as Excel files in dedicated result folders, which are also grouped by channel and region version.

Each folder in the system corresponds to a specific stage in the processing pipeline. For example, separate folders exist for filtered signals, frequency region files, chunked parts, NCD results, and classification summaries. This folder structure supports reproducibility, traceability, and flexibility, making it easy to rerun or inspect any part of the analysis. The entire pipeline was designed to handle large-scale comparisons across 121 participants and 19 channels in a memory-aware and organized manner.

4. Research and Development Process

The research and development process began with organizing and preparing the EEG data files for analysis. Each file, representing one participant, was stored in CSV format and contained signals recorded from 19 standard EEG channels. The first step was to clean the data by applying a bandpass filter ranging from 1 to 40 Hz. This filtering step was essential to eliminate low-frequency drifts and high-frequency noise, thereby focusing on the meaningful components of the signal.

Following the filtering phase, we extracted the five standard EEG frequency bands—Delta, Theta, Alpha, Beta, and Gamma—by segmenting the signals using sliding windows of different durations. Each version (5, 6, 8, and 10 seconds) allowed us to analyze brain activity at varying temporal resolutions. Once the desired frequency region was isolated, we converted the signal into a string-like sequence, divided it into fixed-length chunks of 1000 characters, and stored the parts in structured subfolders organized by channel and participant.

The core of the development process centered around computing the Normalized Compression Distance (NCD) between each part of a participant's signal and the parts of all other participants, using the ZB2 compression algorithm. This step required iterating over large numbers of comparisons for each channel and version. To manage this scale, we built a modular set of Python scripts in Google Colab and organized all intermediate and final results in Google Drive for clarity and reusability.

For each comparison, we calculated the minimum, average, and median NCD values, which were then summarized and saved in Excel files. The median value was used as the main similarity indicator, and final classification was based on comparing these values

against those from known ADHD and control participants. All steps were executed separately for each channel and version, allowing us to analyze classification accuracy in detail.

5. Tools and Environment

The development of the system was carried out entirely in Python using the Google Colab environment, which provided sufficient computational resources and seamless integration with Google Drive for file management. Google Colab's cloud-based execution environment was especially useful in handling the large number of comparisons and managing memory limitations during processing. To structure the EEG data and manage the different processing stages, we designed a consistent folder hierarchy on Google Drive, separating filtered signals, segmented parts, NCD results, and classification summaries by channel and version.

The primary algorithmic component used for similarity measurement was the ZB2 compression algorithm, which was integrated into our Python scripts to calculate the Normalized Compression Distance (NCD). Additional libraries such as pandas, numpy, and os were used for data manipulation, numerical processing, and directory management. The results were saved and summarized using the openpyxl and xlsxwriter libraries, allowing us to export tables and classification values to Excel files for interpretation.

No external client or user interface was required during the development, as the system was intended for internal research use. However, the modularity of our scripts and the structured organization of the outputs make the system easy to adapt for other datasets or use cases. Overall, the combination of cloud computing, compression-based logic, and structured Python development provided an efficient and reproducible workflow for large-scale EEG signal comparison and classification.

6. Challenges and Solutions

Throughout the development of the project, we encountered several challenges that required thoughtful planning, adjustments, and creative solutions. One of the initial challenges involved managing the large number of EEG files and organizing them across different processing stages. With 121 participants, 19 EEG channels, multiple region versions, and hundreds of signal comparisons per channel, the volume of data quickly became substantial. To address this, we created a clearly structured folder hierarchy within Google Drive, categorizing data by participant group, channel, frequency region, and version. This organization enabled efficient access, reduced confusion, and supported repeatability.

Another significant challenge arose from the computational cost of performing NCD comparisons across all pairs of signal segments. Each EEG signal was divided into numerous parts, and the requirement to compare each part with all parts of another participant led to a large number of compression operations. Since compression using the ZB2 algorithm is inherently time-consuming, we needed to design our system to process data in manageable batches, limit unnecessary recomputation, and store intermediate results whenever possible. We optimized our code to perform only essential comparisons, used progress

indicators for tracking execution, and monitored memory usage within the Google Colab environment.

An analytical challenge we faced was selecting the appropriate statistical measure to summarize the NCD values across all segment comparisons. After testing the use of minimum, average, and median values, we observed that the median value offered the most stable and representative results, as it was less sensitive to outlier comparisons that occasionally occurred due to noise or compression artifacts. Choosing the median as our classification metric improved the reliability and consistency of our accuracy measurements.

Finally, working across multiple brainwave frequency regions and time window versions introduced additional complexity. We needed to ensure that all data transformations were consistently applied across different configurations. To manage this, we built flexible functions and reused code components wherever possible. This modular approach reduced errors, streamlined development, and allowed us to adapt the pipeline to multiple scenarios without rewriting major sections of code.

7. Results and Conclusions

To evaluate the effectiveness of our classification approach, we applied three different statistical summarization methods on the NCD values computed between EEG segments of all participants: average (AVG), minimum (MIN), and median (MEDIAN). Each method was tested across five different signal segmentation versions: 2s, 5s, 6s, 8s, and 10s. Below we present the classification performance and observations for each method separately, followed by a general conclusion.

- **Results Using the Average (AVG) Method**

  Using the average NCD values for classification yielded moderate results. As shown in Figure 3, the distribution of scores across participants indicates a spread of performance, with a few participants achieving relatively high classification accuracy (over 60%). However, the method's sensitivity to outlier values in the similarity scores resulted in inconsistent classification performance, particularly for shorter segmentation windows (2s and 5s). The best accuracy achieved using the average method peaked around 72% in a few cases, but overall, this method was outperformed by the median.

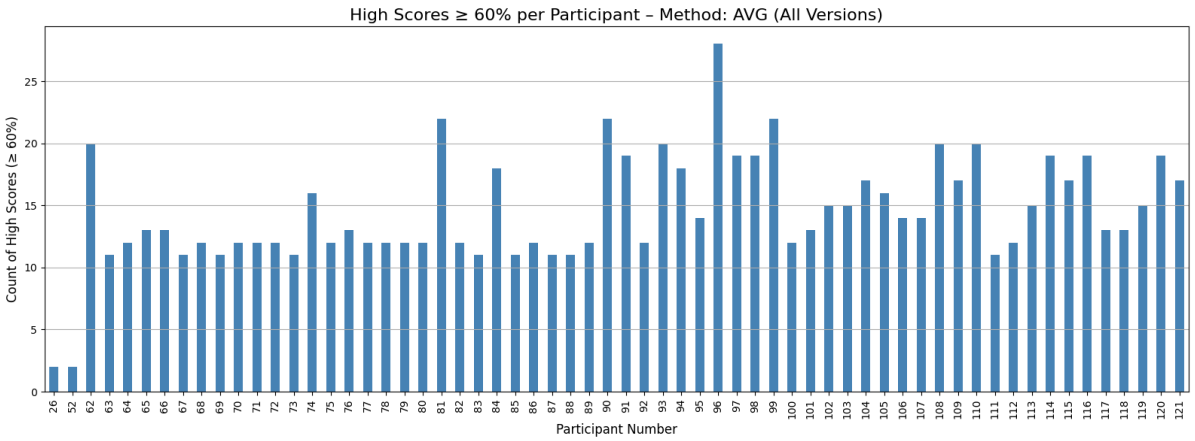

Figure 3 *shows the classification performance for each participant using the Average (AVG) method across all tested segmentation versions (e.g., 5s, 6s, 8s, 10s).*

- *The X-axis represents the participant numbers (1–121), where participants 1–61 belong to the ADHD group, and 62–121 to the control group.*
- *The Y-axis represents the number of classification attempts (out of all versions and channels) where the participant achieved a score ≥ 60%, which we define as a high score.*

- **Results Using the Minimum (MIN) Method**

The minimum-based method showed poor classification performance overall. As evident in Figures 4 and 5, very few participants exceeded the 60% classification threshold. For instance, in the 5s version (Figure 4), only four participants achieved a high score, and in other versions (6s and 8s), classification was nearly random, with little to no participants classified accurately. This method appears to be highly influenced by noise or occasional low-distance matches, which do not reflect meaningful similarity.

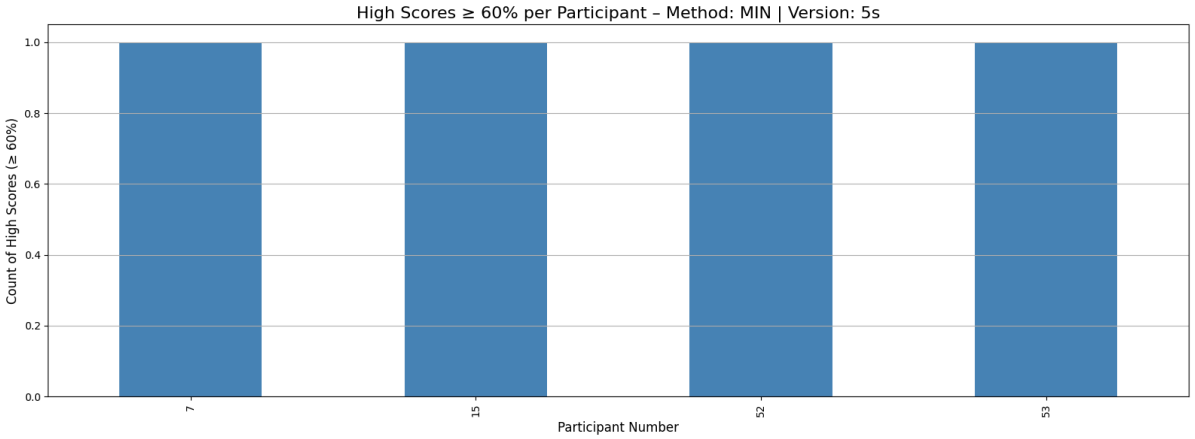

Figure 4 *presents the classification performance using the Minimum (MIN) method with a 5-second segmentation window.*

- *The X-axis shows the participant numbers.*
- *The Y-axis indicates the number of classification scores that reached or exceeded 60% accuracy.*

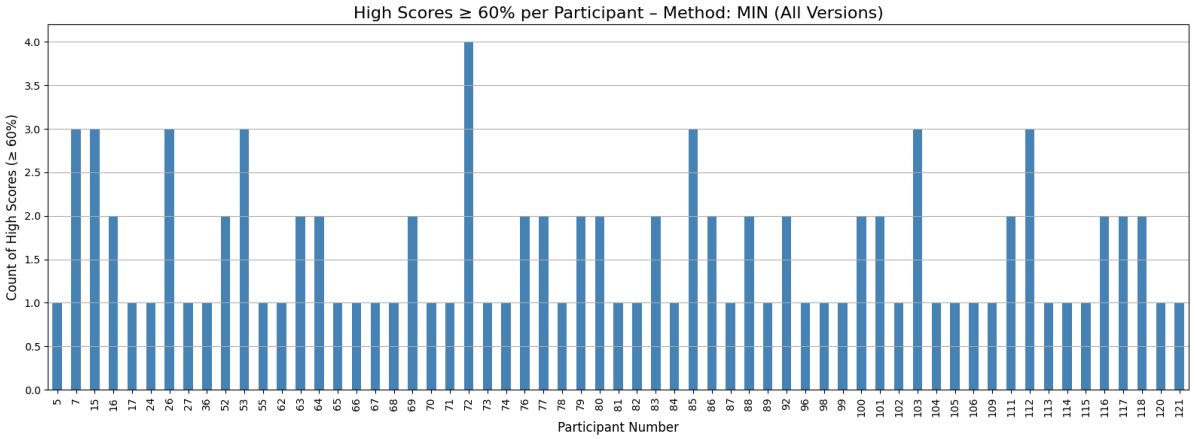

Figure 5 *displays the aggregated classification results using the Minimum (MIN) method across all segmentation versions.*

- *The X-axis represents the participant numbers.*
- *The Y-axis indicates the number of times each participant achieved a classification accuracy of ≥ 60%, which we consider a high score.*

- **Results Using the Median (MEDIAN) Method**

The most stable and accurate classification performance was observed using the median method. As shown in Figure 6, the median filtering of NCD values effectively reduces the effect of outliers and retains central tendencies that better represent participant signal similarities.

When analyzing each version separately, the best performing window was the 10-second version (Figure 7), with many participants achieving 60–80% accuracy. Across all segmentation versions, the overall performance using the median method produced significantly higher classification rates than the average and minimum-based methods.

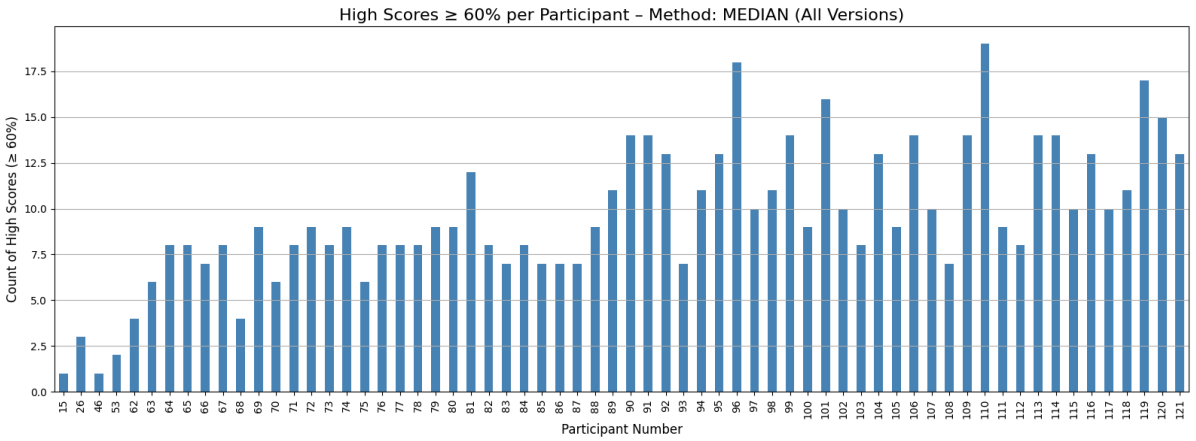

Figure 6 *illustrates the classification performance using the Median (MEDIAN) method across all segmentation versions.*

- *The X-axis represents the participant numbers (1–121).*
- *The Y-axis shows how many classification attempts achieved a score ≥ 60%.*

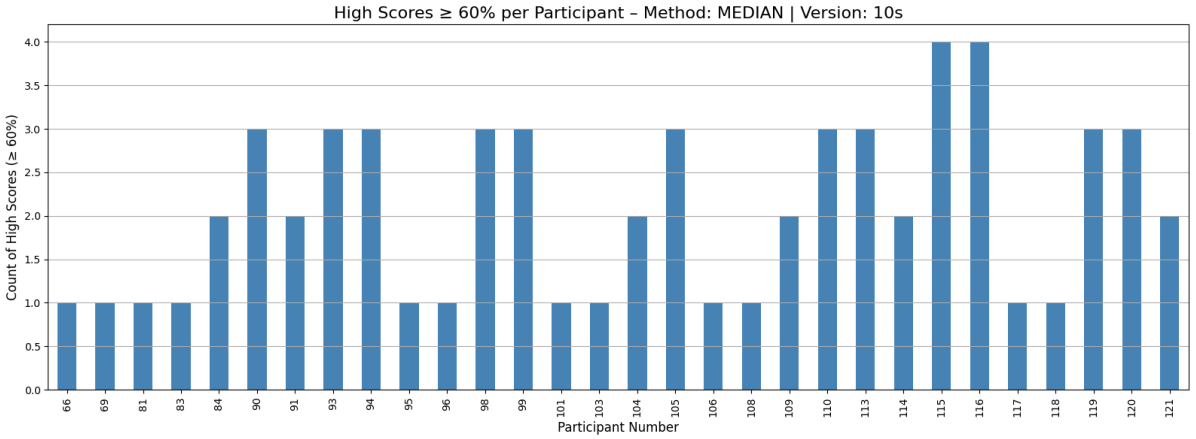

Figure 7 *presents the classification results using the Median (MEDIAN) method with a 10-second segmentation window.*

- *The X-axis shows the participant numbers.*
- *The Y-axis indicates how many times each participant achieved a classification score of ≥ 60%.*

- **Summary of Key Results**

Based on our analysis of classification results with accuracy scores ≥ 60%, the average (AVG) method achieved the highest number of successful classifications, with 899 high-scoring results, followed by the median method with 613, and the minimum method with only 97. These numbers reflect how often each method reached at least 60% classification accuracy, as recorded in the results we analyzed.

Among the five segmentation durations, the 2-second window produced the highest number of successful classifications (565), followed by 8s (320), 10s (248), 5s (238), and 6s (238). These findings suggest that shorter EEG segments may contain more useful signal characteristics for classification using NCD-based similarity.

The minimum method showed the lowest number of high scores and did not perform well in this classification setting. The average method, while producing the most high-score results, also showed variability across channels and participants. The median method provided a different summarization strategy that emphasizes the middle of the similarity score distribution.

Overall, the results demonstrate that using compression-based similarity measures like NCD, combined with appropriate signal segmentation and statistical aggregation, can effectively distinguish between ADHD and control participants. This approach provides a model-free, reproducible framework for analyzing EEG signals using structural similarity.

8. Reflection and Lessons Learned

Throughout the course of the project, we encountered a range of technical and analytical challenges that significantly shaped our understanding of EEG signal processing and compression-based classification. One of the primary lessons learned was the importance of structured data organization. Given the large number of participants, segmentation versions, frequency regions, and methods, maintaining a consistent folder hierarchy and naming convention was essential for reproducibility and ease of debugging.

We also gained valuable insights into the limitations of different statistical summarization methods. Although we initially assumed the minimum value might capture the strongest similarity signal, it became evident that this approach was too sensitive to noise and did not reliably reflect actual classification trends. Conversely, the average and median provided more interpretable and stable summaries, helping us make informed decisions about which signals were truly representative.

Working with compression-based techniques like NCD required careful attention to preprocessing. Ensuring that signals were properly filtered and segmented prior to conversion and compression had a substantial effect on final outcomes. We also learned the importance of tracking intermediate results (e.g., saving part files, NCD values, and summary Excel sheets) to support both analysis and visualization later in the project.

In retrospect, automating more of the processing pipeline—particularly the comparison and classification steps—could have improved efficiency. However, the modular, script-based design we implemented enabled flexible testing across various configurations and was well-suited to exploratory research. Overall, the project provided a deep, hands-on experience in signal processing, similarity measurement, and structured experimentation, all of which are transferable to future research and real-world applications.

References

[1] Cilibrasi, R., & Vitányi, P. M. (2005). Clustering by compression. *IEEE Transactions on Information theory*, *51*(4), 1523-1545.

[2] Cilibrasi, R., Vitanyi, P., & De Wolf, R. (2004, September). Algorithmic clustering of music. In *Proceedings of the Fourth International Conference onWeb Delivering of Music, 2004. EDELMUSIC 2004.* (pp. 110-117). IEEE.

[3] Ziv, J., & Lempel, A. (2003). A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, *23*(3), 337-343.

[4] https://ieee-dataport.org/open-access/eeg-data-adhd-control-children

[5] Python Software Foundation. (2023). *Python Language Reference, version 3.10*. https://www.python.org/

[6] Matplotlib Development Team. (2023). *Matplotlib: Visualization with Python*. https://matplotlib.org/