By Jad Taha & Mohamed Khateeb
Supervisors: Dr. Samah Idres Ghazawy, Dr. Anat Dahan

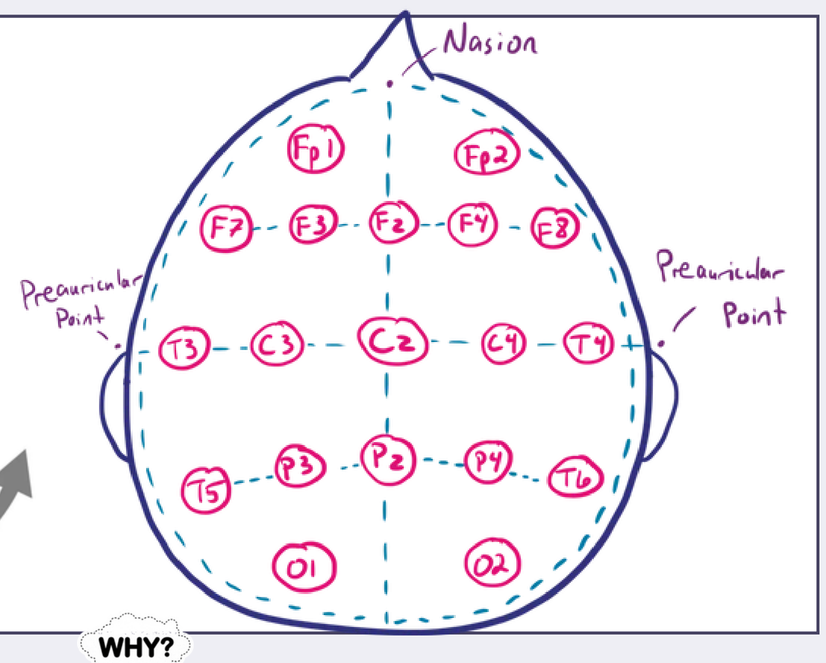
Capstone Project Phase B
EEG Classification Using Normalized Compression Distance for ADHD Diagnosis

BRAUDE College of Engineering, Karmiel

Introduction

This project presents a method for classifying EEG recordings of children with and without ADHD using text compression techniques. Instead of traditional machine learning, we rely on the Normalized Compression Distance (NCD) to measure the similarity between brainwave signals. By filtering, segmenting, and transforming EEG data into symbolic sequences, we compare participants based on how well their signals compress together—offering an interpretable, feature-free approach to EEG classification.

The 10–20 system is a standard method for placing EEG electrodes on the scalp to record brain activity. This figure shows how electrodes are positioned based on head landmarks to cover different brain regions like frontal (F), central (C), and occipital (O).



WHY?

Dataset

- EEG data from 121 children (61 ADHD, 60 Control).
- Recorded using 19 channels at 128 Hz sampling rate.
- Each file contains time-series brain activity per participant.

What is ADHD? What is EEG?

ADHD: “A neurodevelopmental disorder affecting attention and behavior.”

EEG: “A non-invasive method to measure electrical brain activity using scalp electrodes.”

Motiviation

- Early ADHD diagnosis is challenging and often subjective.
- EEG offers an objective way to study brain activity differences.
- We explore a simple, interpretable method without relying on AI.

EEG Recordings

Input: Raw EEG data from ADHD and Control groups

Device: 10–20 system cap, multiple channels

Note: 5–10 seconds of brain signal data per segment

Preprocessing

- Bandpass filtered EEG signals(1–40 Hz) to remove noise.
- Segmented signals into time windows (2s–10s).
- Converted segments to text for compression analysis.

Frequency Transformation

- Converted segments to text for compression analysis
- Represented brainwave patterns as symbolic strings.
- Prepared data for NCD-based similarity comparison.

Comparison

- Computed pairwise similarity using Normalized Compression Distance (NCD).
- Split each string into fixed-length parts (1000 characters)
- Compared every part with every part to another participant using NCD across 19 channels.

NCD Calculation

- Calculated NCD between pairs of EEG sequences using compression.
- Used BZ2 to estimate compressed sizes.
- Lower NCD indicates higher similarity between participants.

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}$$

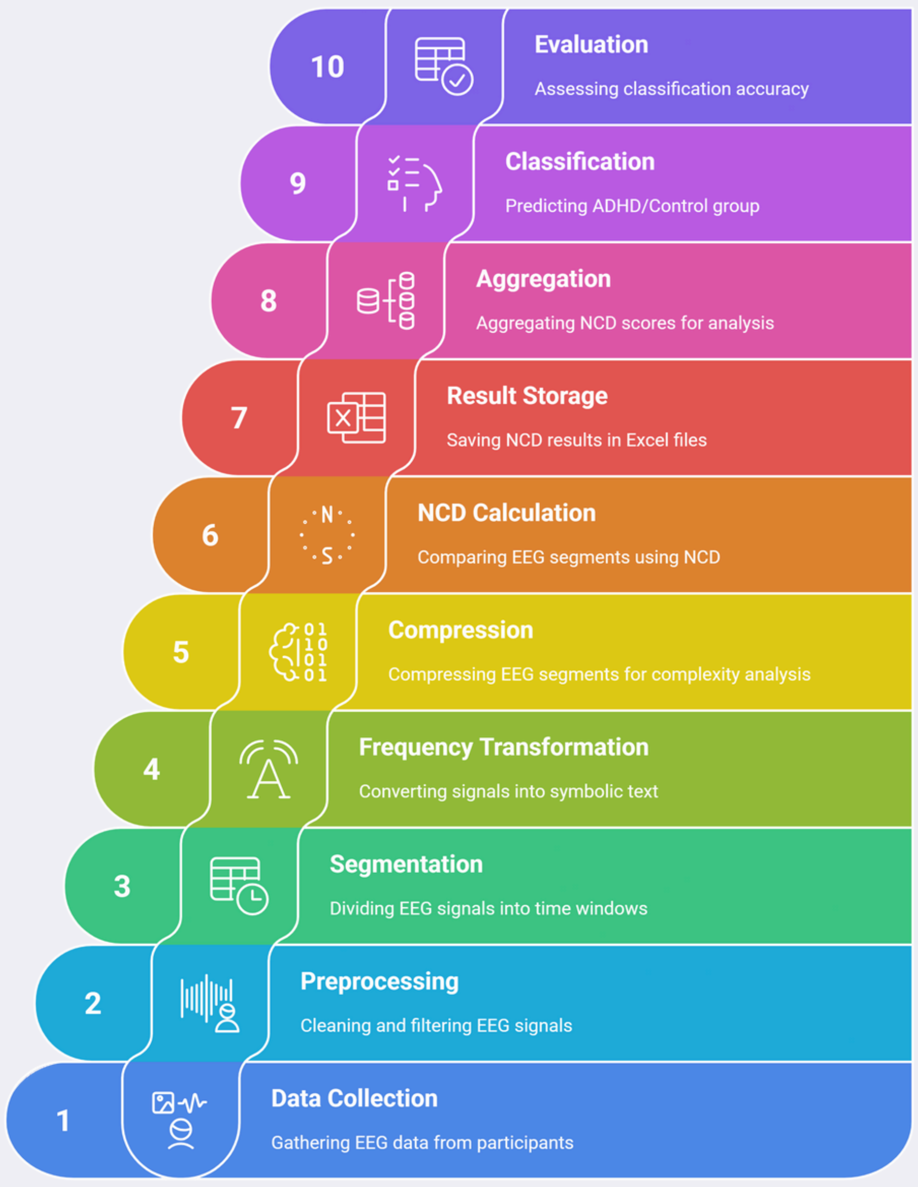
Classification Proccess

- For each comparison file, we computed: Average NCD value, Median NCD value, Minimum NCD value.
- Classified each comparison using the median NCD as a threshold.
- Accuracy was calculated based on how many predict group matched the true labels.

Compression

- Used BZ2 to compress EEG sequences.
- Compression size reflects the complexity of brain signals.

Flow Architecture



10 Evaluation: Assessing classification accuracy

9 Classification: Predicting ADHD/Control group

8 Aggregation: Aggregating NCD scores for analysis

7 Result Storage: Saving NCD results in Excel files

6 NCD Calculation: Comparing EEG segments using NCD

5 Compression: Compressing EEG segments for complexity analysis

4 Frequency Transformation: Converting signals into symbolic text

3 Segmentation: Dividing EEG signals into time windows

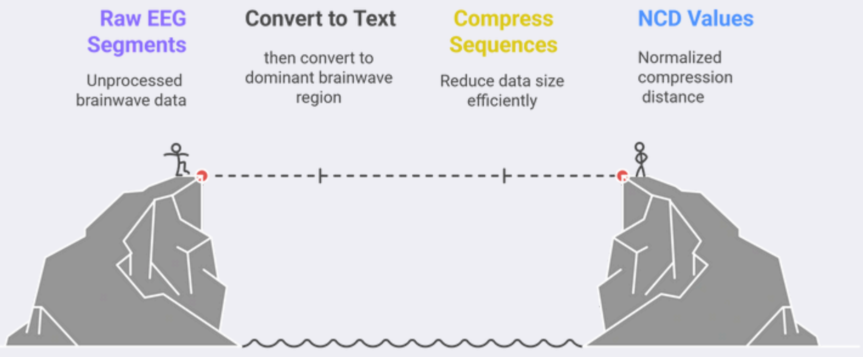
2 Preprocessing: Cleaning and filtering EEG signals

1 Data Collection: Gathering EEG data from participants

Evaluation

- Assessed classification accuracy across all participants.
- Compared performance across segment lengths and aggregation methods.
- Minimum-based strategy showed the highest accuracy overall.

Compression Flow



Raw EEG Segments: Unprocessed brainwave data

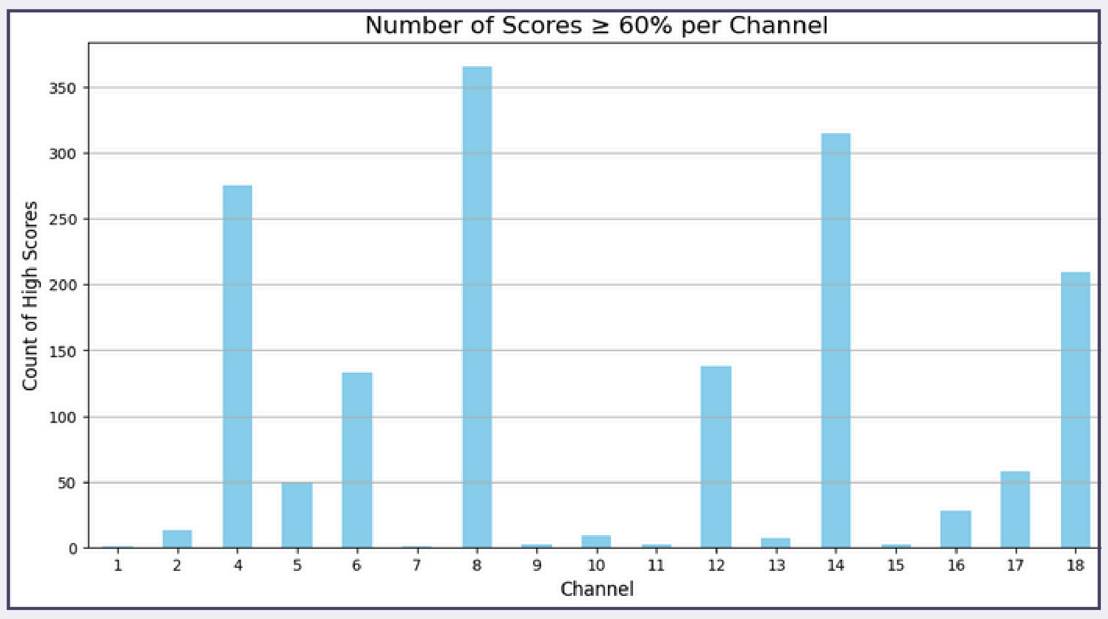
Convert to Text: then convert to dominant brainwave region

Compress Sequences: Reduce data size efficiently

NCD Values: Normalized compression distance

Best Score for Single participant

Method	Average aggregation was used.
Segment Length	Segments were 8 seconds long.
Participant ID	The participant's ID was 110.
Channel	Channel 14 was used for data.

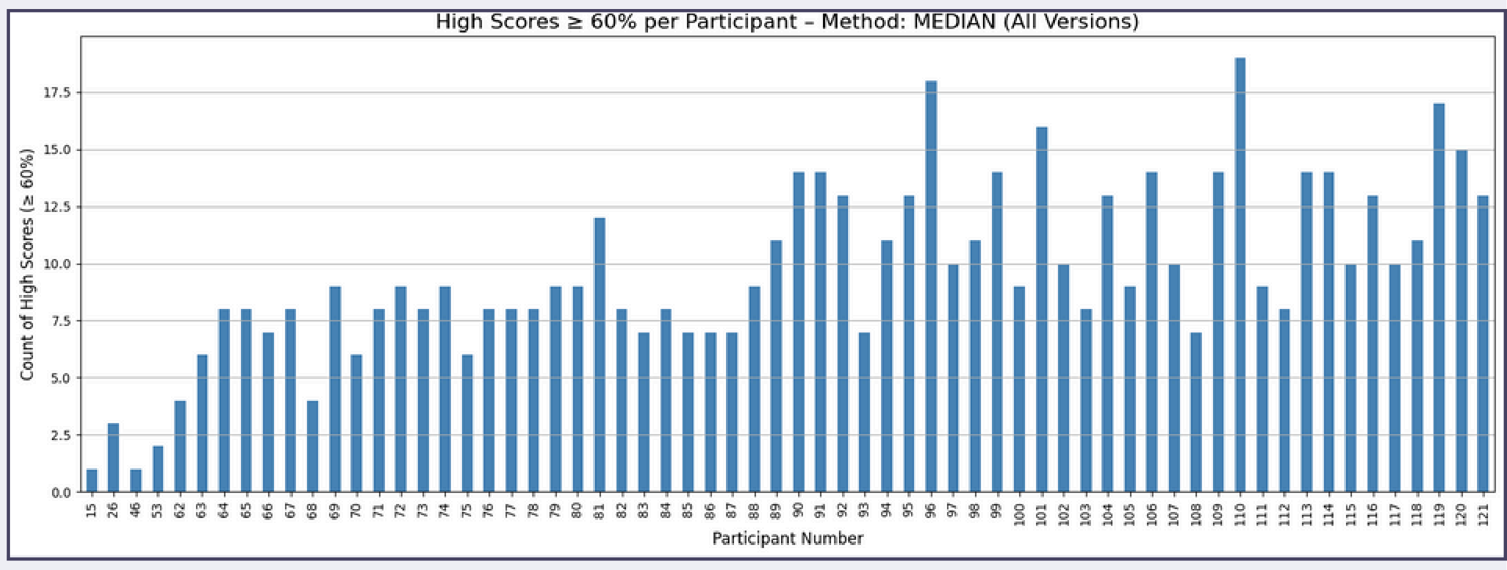


Insights and results

- Minimum-based aggregation gave the highest classification accuracy.
- the shorter time windows (2sec) were more effiictive for NCD based classification.
- Compression effectively captured meaningful differences between ADHD and control signals.
- NCD proved to be a strong, interpretable alternative to machine learning approaches.

Conclusion

This project introduced a simple, interpretable approach for EEG-based ADHD classification using compression techniques. By comparing symbolic representations of EEG signals with Normalized Compression Distance (NCD), we achieved meaningful accuracy without machine learning or feature extraction. The minimum-based method proved most consistent, while the highest individual accuracy was reached using the average method on 8-second segments. These results highlight the potential of compression-based similarity in neurological signal analysis.


High Scores ≥ 60% per Participant – Method: MEDIAN (All Versions)
Count of High Scores (≥ 60%)
Participant Number