Software Engineering Department
Braude College

User Guide

EEG Classification Using Text Compression

25-1-R-2

By:- Mohammad khateeb
Jad taha

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

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

- **Purpose**

  This user guide provides clear operational instructions for running the EEG classification pipeline using Normalized Compression Distance (NCD) with ZB2 compression. The guide walks the user through the key steps required to prepare data, execute the classification workflow, and analyze the output in Google Colab.

- **System Overview**

  The system performs classification of EEG signals to distinguish between ADHD and control participants. It relies on text-based similarity measures using compression techniques rather than machine learning. EEG recordings are preprocessed, segmented, transformed into symbolic form, compared using NCD, and then classified based on similarity statistics.

- **Requirements**
  - Google Colab (recommended: Pro with high RAM)
  - Python 3.x (via Colab)
  - Internet connection
  - Google Drive access with the project folders and files
  - Required Python packages (installed within the notebook):
    - numpy, pandas, matplotlib, scipy, zlib, os, glob, tqdm, xlsxwriter

- **Folder Location**

  Your project should be located in:
  /content/drive/MyDrive/finalProject/

  Ensure the following subdirectories exist:
  - /adhdcsv/ - Original EEG CSV files for ADHD group
  - /controlcsv/ - Original EEG CSV files for Control group
  - /filteredadhdcsv/ - Filtered EEG files (1–40 Hz) for ADHD group
  - /filteredcontrolcsv/ - Filtered EEG files (1–40 Hz) for Control group
  - /brainwave_sequence_2s/, /brainwave_sequence_5s/, /brainwave_sequence_6s/, /brainwave_sequence_8s/, /brainwave_sequence_10s/ – Brainwave region sequences extracted using different time window sizes
  - /parts_ncd_2s/, /parts_ncd_5s/, /parts_ncd_6s/, /parts_ncd_8s/, /parts_ncd_10s/ – Contains the computed NCD values for each part-to-part comparison between participants
  - /min/, /avg/, /median/ – Excel files summarizing the minimum, average, and median NCD values per participant pair
  - /classification_based_on_median/ – Contains per-participant classification results for each method (min, avg, median) using median-based group assignment
  - /high_scores_summary/ – Includes Excel files and graphs showing all classification scores ≥ 60% for each participant and method

- **Files Provided**
  - finalProject2.ipynb – The main notebook for signal processing, part-to-part NCD computation, and classification analysis.
  - /adhdcsv/ – Contains the raw EEG CSV files for ADHD group participants.
  - /controlcsv/ – Contains the raw EEG CSV files for Control group participants.
  - ZB2 compression module – integrated directly within the notebook for computing NCD values.

- **Operating Instructions**
  1. **Open Google Colab** and upload finalProject2.ipynb.

  2. **Mount Google Drive and Install Dependencies**
     ➤ Go to the cell titled:
     #Install libraries, Import modules, Mount Google Drive
     → Run this cell to set up the environment and link to your Drive.

  3. **Verify Folder Structure**
     Ensure the following directories exist in your Google Drive under /finalProject/:
     - /adhdcsv/, /controlcsv/
     - /filteredadhdcsv/, /filteredcontrolcsv/
     - /brainwave_sequence_2s/, /brainwave_sequence_5s/, /brainwave_sequence_6s/, /brainwave_sequence_8s/, /brainwave_sequence_10s/
     - /parts_ncd_2s/, ..., /parts_ncd_10s/
     - /classification_based_on_median/, /high_scores_summary/

  4. **Set Main Project Path**
     ➤ Run the cell titled:
     #Main project folder.
     → Confirm the path is correct and accessible (e.g., /content/drive/MyDrive/finalProject/).

  5. **Filter EEG Signals (1–40 Hz)**
     ➤ Section:
     # **preprocessing functions**
     ➤ Cell:
     # Bandpass filter settings
     → Run all related cells to clean and save filtered data into: /filteredadhdcsv/ and /filteredcontrolcsv/

  6. **Extract Brainwave Regions (per time window)**
     ➤ Section:
     Extracting Dominant EEG Band Sequences from Filtered Signals Using a Sliding Window Approach
     → Run all 5 cells that set output_path for each version (2s, 5s, 6s, 8s, 10s)
     → This will create the /brainwave_sequence_Xs/ folders

7. **Compute NCD Between Parts**

➤ Section:

NCD Funtcion

→ Run the NCD functions for part-to-part comparisons

➤ Then go to the section:

dividing to parts

→ Run all cells to divide the brainwave sequences into 1000-character parts and apply NCD

8. **Calculate min / avg / median NCD per participant**

➤ **Sections to run:**

- # Classification using the parts and average
- # Classification using the parts and median
- # Classification using the parts and min

→ These cells generate participant-to-participant scores and save them to Excel

9. **Final Classification Based on Median Method**

➤ Section:

Classification based on median for (avg, median, min)

→ Run all cells to generate classification decisions for each method

→ Output is saved in /classification_based_on_median/

10. **Generate Final Graphs and Visualizations**

➤ Section:

graphs

→ Run all cells to produce bar charts and summaries

→ Output is saved in /high_scores_summary/

- **Outputs**
  - **Filter EEG Signals (1–40 Hz)**

    After running the bandpass filter step, the system generates filtered EEG signals for each participant. These signals are stored as .csv files, each named in the format:

    filtered_vXp.csv where X represents the participant number (e.g., filtered_v121p.csv for participant 121).

    These filtered files retain all 19 EEG channels and contain only the frequency range between 1 Hz and 40 Hz, effectively removing noise and irrelevant components from the original signal.

0	1	2	3	4	5	6	7	8	9
-54.96700626	-33.56597751	-31.197707	-16.63762478	-26.28492483	-8.060527731	-11.24045243	-11.07164807	13.26316619	-11.1960839
111.1115492	101.6369078	98.47785666	214.2435085	194.1842892	119.0121704	271.5053051	161.4908808	-1280.58743	103.2290195
31.15888549	139.3347184	73.85024228	219.9088136	196.2766757	113.2837306	377.7642766	196.1989342	-979.1412901	124.7301643
-78.52269502	154.5028227	65.87269328	207.4249586	182.0456641	100.8617912	430.0343244	196.8278191	-494.2441632	126.2696137
-55.82664876	178.2007528	113.541634	281.4702117	244.0799592	136.693116	516.6757506	211.0608591	-691.1644866	173.7660584
-5.298812493	201.3008147	85.84653011	301.1664347	274.2979164	163.4679365	602.2034224	225.5640005	-675.3121471	243.5084942
40.49173068	271.8390522	45.81076764	275.9581434	320.3198897	226.2903921	685.7536412	304.3280365	-417.6148328	294.2158733
92.09406956	331.4308193	64.19555927	274.4605544	410.7993439	329.2131756	721.2112575	393.3667	-388.4464172	291.3413398
113.3730637	293.7178279	67.69897757	274.8088889	476.8184467	400.6982172	680.2642014	391.0739824	-387.5768125	255.3780904
139.3830216	261.4477538	77.92022933	307.8231912	553.92204	467.9128597	639.4412128	367.9683816	-322.0289311	232.8041719
211.271669	339.5756252	176.8815701	404.7038792	656.4171801	557.1247113	617.3973336	385.0974737	-341.2553904	205.0700798
276.4926237	439.6954408	286.7220984	482.2354503	675.3488429	579.6164597	568.8797971	363.0594776	-374.2344085	152.8687467
284.6922945	470.4167898	272.5167355	473.3737018	583.8665627	502.4640081	511.5112752	276.8856136	-363.3278092	123.2254165
262.6837709	477.7098673	163.9728057	425.1448754	480.7114749	429.7362188	474.5583439	210.8062015	-385.7381486	144.4751869
258.0224135	505.9075284	103.2574079	397.017543	415.6541133	420.9683741	452.1255052	208.6753666	-401.8914179	189.3291826

The screenshot shows a filtered CSV file, with rows representing time samples and columns labeled 0 to 18, each corresponding to one of the 19 EEG channels. The numeric values represent the filtered brainwave amplitudes per channel at each time step.

- **Extract Dominant Brainwave Regions**

  When you run the dominant region extraction step, the system generates a .txt file for each participant and each EEG channel, representing the dominant frequency region (e.g., Delta, Theta, Alpha, Beta, Gamma) in each sliding window of the signal.

  Each output file is named as:
  vXp_channelY.txt where X is the participant number and Y is the channel index (e.g., v121p_channel18.txt).

  These files contain sequences of characters like DDDDDD..., TTTT..., or BBBB..., where each character stands for the dominant frequency region within a specific time window.

```
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDTTTT
DDDDDDDDDDDDDDDDDDDTTTDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDTTTTDDDDTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDTTTTTTTTDDDDDDDD
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
AAAATTTTTTTTTTTTTAAADDDDDDDDDDDDDDDDDTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
------------------------------------------------------------------------------------
```

  The screenshot shows a dominant region .txt file for one participant-channel pair. Each line is a symbolic representation of

brainwave activity (e.g., 'D' for Delta, 'T' for Theta, 'B' for Beta), capturing the EEG structure in a simplified, compression-friendly form.

- **Compute NCD for Each Part-to-Part**

  When you run this step, the system calculates the Normalized Compression Distance (NCD) between each pair of parts from two participants' signals for a specific channel. These part-to-part comparisons produce a CSV file that lists every possible match between segments, including the computed similarity value (NCD).

  Each output file is named:
  vXp_vs_vYp_channelZ.csv

  Where:
  - X is the first participant number
  - Y is the second participant number
  - Z is the EEG channel index

  Each row in the file corresponds to:
  - Part #X and its symbolic sequence
  - Part #Y and its symbolic sequence
  - The NCD value between the two parts

A	B	C	D	E
Part #120p	Sequence 120p	Part #121p	Sequence 121p	NCD Value
1	DDDDDDDDDDDDDDDDDDDD	1	DDDDDDDDDDDDDDDDDDDD	0.3582089552
1	DDDDDDDDDDDDDDDDDDDD	2	DDDDDDDDDDDDDDDDDDDD	0.3582089552
1	DDDDDDDDDDDDDDDDDDDD	3	TTTTTTTTTTTTTTTTTTTTTTTT	0.3432835821
1	DDDDDDDDDDDDDDDDDDDD	4	TTTDDDDDDDDDDDDDDDDTT	0.44
1	DDDDDDDDDDDDDDDDDDDD	5	DDDDDDDDDDDDDDDDDDDD	0.3529411765
1	DDDDDDDDDDDDDDDDDDDD	6	DDDDDDDDDDDDDDDDDDDD	0.3880597015
1	DDDDDDDDDDDDDDDDDDDD	7	DDDDDDDDDDDDDDDDDDDD	0.3432835821
1	DDDDDDDDDDDDDDDDDDDD	8	DDDDDDDDDDDDDDDDDDDD	0.4
1	DDDDDDDDDDDDDDDDDDDD	9	DTTTTTTTTTDDDDDTTTTTTTTT	0.328358209
1	DDDDDDDDDDDDDDDDDDDD	10	DDDDDDDDDDDDDDDDDDDD	0.2985074627
1	DDDDDDDDDDDDDDDDDDDD	11	DDTTTTTTTTDDDDDDDDDDDD	0.328358209
1	DDDDDDDDDDDDDDDDDDDD	12	DDDDDDDDDDDDDDDDDDDD	0.3880597015
2	DDDDDDDDDDDDDDDDDDDD	1	DDDDDDDDDDDDDDDDDDDD	0.4242424242
2	DDDDDDDDDDDDDDDDDDDD	2	DDDDDDDDDDDDDDDDDDDD	0.3939393939
2	DDDDDDDDDDDDDDDDDDDD	3	TTTTTTTTTTTTTTTTTTTTTTTT	0.3939393939
2	DDDDDDDDDDDDDDDDDDDD	4	TTTDDDDDDDDDDDDDDDDTT	0.4933333333
2	DDDDDDDDDDDDDDDDDDDD	5	DDDDDDDDDDDDDDDDDDDD	0.4558823529
2	DDDDDDDDDDDDDDDDDDDD	6	DDDDDDDDDDDDDDDDDDDD	0.4393939394
2	DDDDDDDDDDDDDDDDDDDD	7	DDDDDDDDDDDDDDDDDDDD	0.3939393939
2	DDDDDDDDDDDDDDDDDDDD	8	DDDDDDDDDDDDDDDDDDDD	0.4571428571
2	DDDDDDDDDDDDDDDDDDDD	9	DTTTTTTTTTDDDDDTTTTTTTTT	0.3939393939
2	DDDDDDDDDDDDDDDDDDDD	10	DDDDDDDDDDDDDDDDDDDD	0.3939393939
2	DDDDDDDDDDDDDDDDDDDD	11	DDTTTTTTTTDDDDDDDDDDDD	0.4090909091
2	DDDDDDDDDDDDDDDDDDDD	12	DDDDDDDDDDDDDDDDDDDD	0.4393939394
3	DDDDDDDDDDDDDDDDDDDD	1	DDDDDDDDDDDDDDDDDDDD	0.3846153846
3	DDDDDDDDDDDDDDDDDDDD	2	DDDDDDDDDDDDDDDDDDDD	0.4
3	DDDDDDDDDDDDDDDDDDDD	3	TTTTTTTTTTTTTTTTTTTTTTTT	0.4
3	DDDDDDDDDDDDDDDDDDDD	4	TTTDDDDDDDDDDDDDDDDTT	0.52
3	DDDDDDDDDDDDDDDDDDDD	5	DDDDDDDDDDDDDDDDDDDD	0.4264705882
3	DDDDDDDDDDDDDDDDDDDD	6	DDDDDDDDDDDDDDDDDDDD	0.3692307692

The image shows an output CSV with symbolic sequences from Participant 120 and Participant 121 for Channel 18 (2-second version). The final column contains the NCD similarity values. These values quantify how structurally similar two sequences are based on compression—lower values imply higher similarity.

- **Calculate min / avg / median NCD per participant**

  When you run this step, the system summarizes the NCD values between a selected participant and all other participants for a specific channel, based on the chosen method: minimum, average, or median. Each output is a .txt file containing:

  - The filename of the original NCD comparison file
  - The computed NCD score using the selected method

  Each file is named and stored in the following structure:/<method>/<Xs>/channelY/participantZ.txt
  Where:

  - method is one of avg, min, or median
  - X is the window size (2s, 5s, 6s, 8s, 10s)
  - Y is the channel number
  - Z is the participant number

```
Participant 121 - Channel 18 (Min-Based)
v76p_vs_v121p_channel18.xlsx: 0.02326
v77p_vs_v121p_channel18.xlsx: 0.02326
v78p_vs_v121p_channel18.xlsx: 0.02326
v79p_vs_v121p_channel18.xlsx: 0.02326
v80p_vs_v121p_channel18.xlsx: 0.02326
v81p_vs_v121p_channel18.xlsx: 0.02326
v82p_vs_v121p_channel18.xlsx: 0.02326
v83p_vs_v121p_channel18.xlsx: 0.02326
v84p_vs_v121p_channel18.xlsx: 0.02326
v85p_vs_v121p_channel18.xlsx: 0.02326
v86p_vs_v121p_channel18.xlsx: 0.02326
v87p_vs_v121p_channel18.xlsx: 0.02326
v88p_vs_v121p_channel18.xlsx: 0.02326
v89p_vs_v121p_channel18.xlsx: 0.02326
v90p_vs_v121p_channel18.xlsx: 0.02326
v91p_vs_v121p_channel18.xlsx: 0.02326
v92p_vs_v121p_channel18.xlsx: 0.02326
v93p_vs_v121p_channel18.xlsx: 0.02326
v94p_vs_v121p_channel18.xlsx: 0.02326
v95p_vs_v121p_channel18.xlsx: 0.02326
v96p_vs_v121p_channel18.xlsx: 0.02326
v97p_vs_v121p_channel18.xlsx: 0.02326
v98p_vs_v121p_channel18.xlsx: 0.02326
v99p_vs_v121p_channel18.xlsx: 0.02326
v100p_vs_v121p_channel18.xlsx: 0.02326
v101p_vs_v121p_channel18.xlsx: 0.02326
v102p_vs_v121p_channel18.xlsx: 0.02326
v103p_vs_v121p_channel18.xlsx: 0.02326
v104p_vs_v121p_channel18.xlsx: 0.02326
v105p_vs_v121p_channel18.xlsx: 0.02326
v106p_vs_v121p_channel18.xlsx: 0.02326
v107p_vs_v121p_channel18.xlsx: 0.02326
v108p_vs_v121p_channel18.xlsx: 0.02326
v109p_vs_v121p_channel18.xlsx: 0.02326
v111p_vs_v121p_channel18.xlsx: 0.02326
```

The image displays the contents of participant_121.txt located under /min/2s/channel18/. It lists NCD scores between participant 121 and other participants using the minimum-based method for each comparison. All values are consistent here (0.02326), highlighting similarity patterns across comparisons.

- **Final classification Based on median method**

When you run this step, the system generates an Excel file for each participant that summarizes their classification scores across all EEG channels using three different methods: average (avg), median, and minimum (min). These scores reflect how well the participant's EEG signals align with others from the same or opposite group, based on the chosen metric.

Each output file follows the format:
participant_X.xlsx , Where X is the participant number.

These files are stored in the folder:
/classification_based_on_median/version/channelY/

	Channel 0	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5	Channel 6	Channel 7	Channel 8	Channel 9
avg	0.49167	0.55	0.55833	0.55833	0.625	0.575	0.60833	0.525	0.575	0.525
median	0.50833	0.50833	0.53333	0.525	0.53333	0.60833	0.55	0.48333	0.60833	0.55
min	0.49167	0.5	0.49167	0.50833	0.49167	0.48333	0.5	0.5	0.51667	0.50833

The image shows a classification summary table from participant_121.xlsx, containing scores across channels 0–18. The rows correspond to the three methods (avg, median, min), and the columns represent EEG channels. Higher values typically indicate stronger classification confidence.

- **High Score Summary & Graphs**

When you run this step, the system generates an Excel file summarizing all classification results where the score is ≥ 60%. The file includes:

- The method used (avg, median, or min)
- The segmentation window size (e.g., 2s, 5s, 10s)
- The participant number
- The EEG channel on which the score was calculated
- The classification score

The goal is to track strong classification results across all conditions and help analyze patterns of performance.

These results are saved in the folder: /high_scores_summary/

In addition to the Excel file, this step generates bar chart graphs visualizing:

- Number of high scores per participant
- Method-wise comparison
- Performance across segmentation versions

A	B	C	D	E	F
score number	version	score	participant	channel	method
1290	8s	0.68333	110	14	avg
32	2s	0.675	63	14	median
58	2s	0.675	65	14	median
301	2s	0.675	89	14	median
429	2s	0.675	106	14	avg
506	2s	0.675	115	8	avg
551	2s	0.675	120	14	avg
1138	8s	0.675	84	14	avg
1166	8s	0.675	90	14	avg
1320	8s	0.675	115	14	avg
1438	10s	0.675	90	8	avg
13	2s	0.65833	26	4	min
53	2s	0.65833	65	14	avg
63	2s	0.65833	66	14	avg
77	2s	0.65833	67	14	median
122	2s	0.65833	72	14	avg
156	2s	0.65833	75	14	avg
247	2s	0.65833	84	14	avg
262	2s	0.65833	85	14	median
273	2s	0.65833	86	14	median
284	2s	0.65833	87	14	median
380	2s	0.65833	100	14	median
387	2s	0.65833	101	14	avg
500	2s	0.65833	114	14	avg
510	2s	0.65833	115	8	median
525	2s	0.65833	117	8	median

This image shows a portion of the high_scores_summary.xlsx file. Each row represents a high-scoring classification instance. You can see the segmentation version, the score, the participant, the EEG channel, and the method used.