Review of: "A Simple Preprocessing Method Enhances Machine Learning Application to EEG Data for Differential Diagnosis of Autism"

Ana Rodrigues[1]

1 Kaunas University of Technology

Potential competing interests: No potential competing interests to declare.

The manuscript “A Simple Preprocessing Method Enhances Machine Learning Application to EEG Data for Differential Diagnosis of Autism” proposes a clustering algorithm based on the minimum spanning tree (MST) to group temporal differences between the 19 EEG channels. The authors hypothesize that children with Autism Spectrum Disorder (ASD) have distinguishable temporal differences from those shown by children with other neuropsychiatric disorders (NSD). The MST algorithm compresses each EEG signal into 38 numbers, where each electrode is associated with a specific number of links. The Manhattan distance matrix between each pair of electrodes is applied to calculate the number of links for each electrode, in which longer distances indicate a more preeminent disconnection, i.e., dissimilarity, between two brain regions. Unsupervised machine learning methods, such as KNN, were then employed to cluster MST-derived features, showing a notable separation between children with ASD and NSD. The authors report a mean accuracy of 93.20%. Since ASD is often misdiagnosed as another NSD, a quantitive screening tool for ASD is of clinical importance. Thus, the manuscript presents research that is relevant to the field.

While the approach proposed in the manuscript is **interesting**, the authors must address many other questions and describe the methodology in more detail so that other studies can recreate it. Moreover, the research design requires substantial improvement. Below, I list some points that require improvement and clarification.

Remarks about the research design

The data cohort comprises 50 children with ASD and 50 with other NPD, both groups with similar age (4–10 years old) and sex (39 boys, 11 girls) distribution. Some of the primary NPD diagnoses included ADHD, mood, and anxiety disorders, which are known frequent comorbidities of ASD. A few questions arise here:

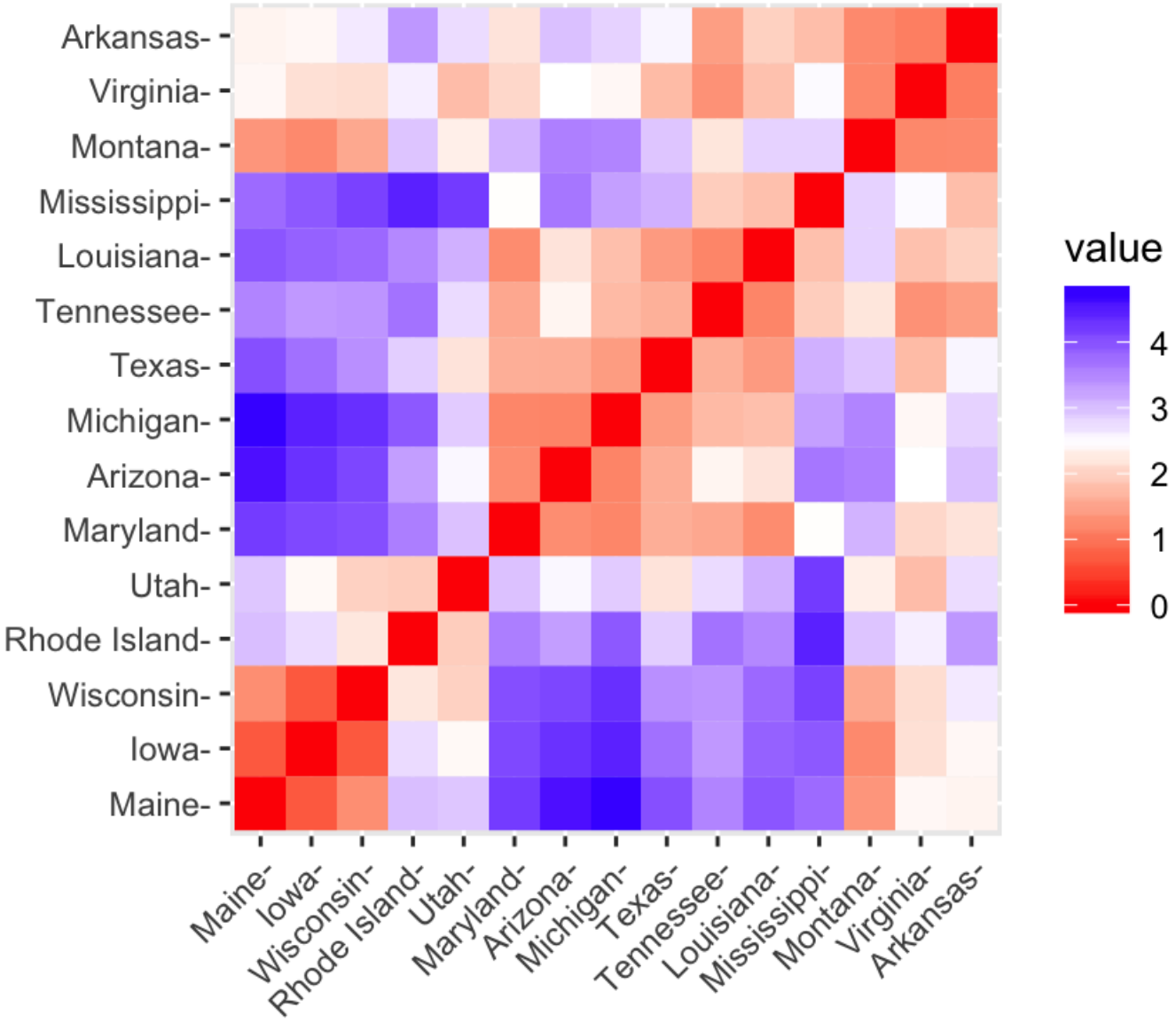
1. How confident are the authors that none of these children also has ASD or vice versa, i.e., that the ASD group presents a comorbid NPD (e.g., ADHD)?
2. What are the symptoms of the ASD group? Or rather, what is the exact description of the ASD group (e.g., how many are non-verbal and have cognitive impairments)? I would imagine that children with conspicuous symptoms of autism would display more noticeable topological EEG differences and, therefore, yield better cluster separation between the groups. Even if this is not the case, the authors must clarify their cohort description.

3. Boys and girls also have distinctive presentations of ASD and other NPDs, such as ADHD, which is why many girls often go undiagnosed. Biological differences between the brains of boys and girls are the speculated reason for such differences in disorder presentation; thus, it would be interesting to include a more detailed analysis of the algorithm's results according to sex. I'd also be wary of stating that the male/female ratios are matched since 78% of the cohort are boys. The authors should expressly state that the training and testing datasets follow the same distribution as the whole database, including in the abstract.

Remarks about the methodology and results

The methodology is written as a school report rather than a scientific paper. Furthermore, the research design requires reconsideration. A few remarks/questions:

1. How exactly is the **time distance** between each pair of electrodes measured (and what units)? The authors could include an illustration of this parameter and add units to Figure 1.
2. The authors label the MTS algorithm as the "*preprocessing* phase." However, preprocessing in digital signal analysis often refers to artifact removal (e.g., filtering) rather than feature extraction, which is what the authors are essentially doing. In fact, the authors state that the EEG signal is artifact-free. Thus, the manuscript would be clearer if this phase were renamed to "EEG feature extraction" or something similar.
3. The introduction also contains an extensive paragraph about artifact removal that is irrelevant to the article, and it would be more productive to describe what has been done on EEG-based detection of ASD and the shortcomings of such studies. For example, the information in Table 2 seems more pertinent to the introduction and methodology sections than the discussion.
4. Why did the authors use the Manhattan distance instead of other metrics? Were other distances tested? Examining the impact of various distance metrics on the algorithm's accuracy would also be a suitable addition to the experimental research results.
5. What is the difference between the KNN algorithm and the Pick and Squash Tracking (PST) algorithm? The reason why the PST algorithm was used (and where) needs to be clarified.
6. Table 1 also shows two rows with KNN. What is the difference between them? Is one of the rows supposed to be the PST algorithm?
7. Why is the split 50% instead of the traditional 70/30 or 80/20? If data shortage is the issue, the authors can use k-fold cross-validation while ensuring a homogeneous distribution between the training and testing datasets. A 50/50 split is more prone to statistical biases and overfitting.
8. What do the authors mean by "ANN" in the training-testing protocol? I am trying to understand why an artificial neural network is necessary for splitting the data.
9. The data visualization also needs substantial improvements. Figure 1 does not contain a caption, nor is it a figure. Dissimilarity matrices are typically visualized with a heatmap:



#Example of a heatmap.

1. Figure 4 is also not a figure but rather a table. I'd recommend presenting it as a horizontal table and adding the corresponding electrode to each number, as stated in the last two lines of page 5.

The authors must be **careful about abbreviations** and include them in the text one time. MST, for instance, appears twice on page 3, but ADHD, KNN, ASD, and ANN do not appear anywhere.

The discussion also could use some improvement. The authors could explain whether there is a plausible biological explanation for such temporal differences in EEGs and highlight the study's limitations.

Final decision

Although I understand that Qeios is not a journal and the manuscript cannot technically be published, I would **not recommend it for acceptance in the current format**. However, I do think the manuscript has potential to become one

suitable for publication if additional experiments and re-writing is done.