Author’s response to reviews

Title: Machine Learning Analysis of Motor Evoked Potential Time Series to Predict Disability Progression in Multiple Sclerosis

Authors:

Jan Yperman (jan_yperman@uhasselt.be; yperman.jan@outlook.com)
Thijs Becker (thijs.becker@uhasselt.be)
Dirk Valkenborg (dirk.valkenborg@uhasselt.be)
Veronica Popescu (veronica.popescu@uhasselt.be)
Niels Hellings (niels.hellings@uhasselt.be)
Bart Van Wijmeersch (bart.vanwijmeersch@uhasselt.be)
Liesbet M Peeters (liesbet.peeters@uhasselt.be)

Version: 3 Date: 13 Dec 2019

Author’s response to reviews:

The authors would like to thank all the reviewers for their valuable comments. We have modified the manuscript to incorporate these. A pdf in which these changes are highlighted is included as supplementary material. We now address all the comments individually:

Miguel García Torres PhD:

COMMENT:
Model evaluation. In this work the models were tested using different training-test splits. In this work the random forest are evaluated using cross validation (Classifier, section 2.2). However I read in page 11 that the authors ran the pipeline using several splitting values. For me it is not clear how the models are tested.

ANSWER:
We have added a more extensive explanation of the cross-validation process in section 2.3, and in the additional file. We replaced Figure 2 with a more comprehensive illustration of the data analysis pipeline. We agree that this was not clear from the manuscript.

COMMENT:
In my opinion, they should use stratified cross-validation with a k=10 or k=5 if it is time-consuming.

ANSWER:
The results obtained by 5-fold stratified cross-validation are too unstable, which is why we opted for our current approach. To illustrate this, we ran the pipeline using this performance measure for 3
different seeds of the splitting algorithm. These are the results:

Seed #1: Literature: 0.724 (0.073)  Literate + extra TS features: 0.757 (0.060)
Seed #2: Literature: 0.727 (0.081)  Literate + extra TS features: 0.741 (0.062)
Seed #3: Literature: 0.733 (0.062)  Literate + extra TS features: 0.752 (0.055)

These results are too unstable, with a difference in score of up to 0.01. 10-fold CV would further halve the test set size, leading to even larger standard deviations. For the final performance of the model we average the performance across 1000 grouped (by patients) stratified shuffle splits, for different ratios of train/test sizes. We have adjusted the explanation in Section 2.3 to clarify this and illustrated it in Figure 2.

For determining the values for the hyperparameters we are currently using grouped k=4 cross-validation. We have clarified this in the explanation of the cross-validation process in section 2.3, and with a better visualisation in Figure 2. (See also our previous response.)

A test on the robustness of the hyperparameter selection has been added in the appendix. The extra experiments we carried out confirmed that the performance is stable for large intervals of the hyperparameter values. This discussion is added in the additional file.

COMMENT:
Table 2 should be replaced since it makes no sense to perform several experiments by changing the percentage split. If the dataset is small, then apply cross-validation. If it is large enough, then holdout (2/3 training and 1/3 test).

ANSWER:
By changing the percentage of training data, one can assess the amount of data necessary to achieve a certain performance, and estimate the room for improvement by including extra data. We therefore believe it is an informative experiment to perform. We added the following text in the paper to clarify this:

"This method also gives information on the necessary dataset size to achieve a certain performance [56, 57] and on how much room for improvement the algorithm has when given more data [58, 59]


COMMENT:
- Feature selection algorithm. Authors claim that the use of Boruta's algorithms performed poorly and, so, applied Mutual Information to filter the features. However, it is not possible to know a priori, the number of features to remove since the perform a ranking, which measures single interaction between a
feature and the class and, so, it does not take into account possible interaction among features. Therefore, I would recommend to apply another techniques more suitable for feature selection:

1) Fast Correlation Based Filter [1]

2) Variable Neighbourhood Search for high dimensional data [2]

Although [1] doesn't either take into account interaction among features, it removes redundant features and able to find good subsets of features.

So, in my opinion the experiments should be repeated applying cross-validation and the feature selection step should be extended and use FCBF and VNS algorithms.


ANSWER:
Besides the 4 other approaches we added after the first review round, we have added a 5th approach based on the “Fast correlation based filter” (FCBF) suggestion. The results are in agreement with our previous experiments: FCBF was not able to beat the pipeline used in our work. These approaches are described in the additional file.

We were interested in applying the approach based on feature grouping, but we did not find software packages that implement this method. Because of time constraints and reproducibility issues, we therefore chose to not implement this method.

We have added a further explanation of why we believe Boruta works so well: in Section 2.3 we added: “Because the Boruta method is based on random forests, and because we use random forests as our classifier, we expect it to be well suited for the feature selection task.”

COMMENT:
José Luis Ayala (Reviewer 1): The authors have improved the paper paying attention to the provided review. Therefore, I recommend the paper for publication.

ANSWER:
We thank the reviewer for his positive feedback.

COMMENT:
Manuel Vázquez-Marrufo (Reviewer 2): NURL-D-19-00203R2
Machine Learning Analysis of Motor Evoked Potential Time Series to Predict Disability Progression in Multiple Sclerosis
After reading the answers from the authors, I still have some concerns about this manuscript:

1. If I have understood well, the main goal of the present study is to improve the possibility of prognostic value using EPTS (specifically MEP) and machine learning.

ANSWER:
This is indeed the main goal. To clarify our main goal, we have rewritten the introduction, and replaced “prognosis” (which is quite general and could mean several things) with the more specific terms “prediction of disability” and “monitoring of disease course”. We hope that the new introduction is less vague and more precise.
COMMENT:
Considering that authors refer to "time series" I expect some intervals of points and not only latency or peak to peak measures. Indeed, in the abstract section authors stated "We perform a machine learning analysis of motor EP that uses the whole time series..". Checking the figure 1 it is not marked or defined intervals that are included in the measurement. I would suggest a more clear definition of what is analyzed in this study and included as data for machine learning algorithms.

ANSWER:
It is clear that we did not explain well enough on what part of the time series the feature extraction is done. In contrast to, e.g., visual evoked potentials or somatosensory evoked potentials, we don’t need to define intervals: latencies are annotated by specialized nurses. The extra time-series features are extracted from the full signal, so no interval needs to be defined.
To make this more clear, we have added a detailed description of the measuring procedure of the motor evoked potentials in a new section (Section 2.1: Measurement protocol).
We have also expanded on the explanation in Section 2.3, in “feature extraction”. In the new version we write (new part in bold):
“
Because each EPTS starts with a large peak at the beginning, an uninformative artifact of the electrophysiological stimulation, the first 70 samples of each EPTS are discarded. A diverse and large set of time series features is extracted from the rest of the EPTS (1850 samples) with the HCTSA package, which automatically calculates around 7700 features from different TS analysis methodologies.
”

COMMENT:
2. About noise, this point is highly relevant. The main concern is if authors have used discrete points of measurement (latency or amplitude of peaks) or collection of data in a specific interval. In the second case, noise could be determinant for the results of the application of any algorithm. Signal-noise ratio is one measurement that can help to value the cleanness of the MEPs. Another possibility is to present a grand average of MEP from all subjects just to see the baseline and the MEP wave compared to that.

ANSWER:
Since this was not clear in the manuscript, we have further clarified in the new version that there is no collection of data in a specific interval: we use the full time series to extract features (see previous answer).
Regarding the impact of noise on the features: the machine learning algorithm should be able to deal with the noise, since this is how the MEP are measured in routine clinical follow-up.

COMMENT:
In any case, specific details of the recording are also missed, for instance, system employed for the study, digitizing rate, time interval used, and so on.

ANSWER:
Specific details have been included in the new section 2.1: measurement protocol.
COMMENT:
3. Lastly, I still found the discussion and conclusions far away to be directly applied in the neurology field. As an example, figures that probably are representing diverse results of the algorithms, they do not help in my opinion to understand the potential application in the MS prognostics. My recommendation to the authors is to make a special effort to explain in a more clear way the benefits of their methods for the neurologist.

ANSWER:
The introduction has been rewritten, to make it shorter and more to the point, and to make the relevance to the prediction of disability progression in MS more clear. We hope that the last part in the discussion also shows the potential benefits for neurologists using EP in clinical practice. We made a small change in the conclusion section, to highlight the usefulness for more certainty about the disease evolution, which is relevant for patients.