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: 4 Date: 30 Jan 2020

Author’s response to reviews:

We thank the reviewers for their comments and continued constructive feedback. In the text below, we tried to address the issues raised as much as possible. We hope that this fully clarifies the concerns (both general as well as specific) raised by the referees. If this is not the case, we are open to have a more direct conversation with the referee (e.g. via Skype), if the editor agrees that this is possible. We believe this can resolve any possible remaining issues in a faster and more effective manner. For convenience we have included a reference pdf where the changes in the manuscript since the previous review are highlighted, called 'Manuscript diff'.

Reply: general issues

Besides the specific points raised by reviewer 2 (see below), there seem to be 2 general clarifications we need to make:

[1] We are not investigating “EEG type” evoked potentials, where the response is measured by electrodes on the brain. The paper by Coppola et al. (2010) that is referred to is about the pre-processing of somatosensory EPs, which is an “EEG type” EP. Similarly, the article by Gonzalez-Rosa et al. (2011) in BMC Neurology (of which the referee is a co-author), is about event-related potentials (ERPs). ERPs are also an “EEG type” measurement. For these “EEG type” evoked potentials, one needs to take into account things such as:
   * averaging several 100s or 1000s of sweeps (to reduce noise)
   * consider signal-to-noise ratios (to quantify how much noise there is)
   * define several different latencies per time series (e.g. N20 and P25)
   * consider time windows to extract these latencies
   * measure a baseline before the stimulus
It is this type of pre-processing description which referee 2 seems to be expecting (by referring to Coppola et al. (2010) for guidance and mentioning ERPs). However, we are studying motor evoked potentials. These are not “EEG type” EPs that are measured on the brain; rather, they are measured in the muscles, and the excitation happens by transcranial magnetic stimulation (TMS). The issues described above do not occur here:

* there are no sweeps, there is only one measurement per magnetic excitation.
* signal-to-noise ratios are not typically calculated for MEPs (at least, we are not aware of a paper where this is done for MEPs)
* there is only one latency: the time it takes for the signal to arrive; no other latencies are defined here

  * since only the latency is needed, and because these are annotated manually by a nurse, there are no time windows to define. The measurement starts the moment the motor cortex is stimulated.
  * the measurement starts at the time of the magnetic stimulation. There is no measurement of a baseline before the stimulus, though there is some time between the stimulation and the muscle response which is included in the measurement. We’ve illustrated this in the new Figure 1.

Our description and included details about the measurement protocol and pre-processing of motor EP is similar to other such descriptions (e.g. Giffroy et al., BMC Neurology 2016; Fuhr et al., Brain 2001; Iodice et al., Journal of the Neurological Sciences 2016).

[2] We use “samples” to mean 2 different things:

1. If a time series is measured, its sampling rate (or frequency of digitization) leads to discrete values that are measured from the (in principle continuous) time series. Each time point that is sampled is referred to as a “sample”.
2. In machine learning, one input-output pair is often referred to as a “sample”. Here, one input-output pair is:
   - input = variables measured from one patient visit (latency, peak-to-peak amplitude, age, …)
   - output = has the patient shown progression of disability after 2 years (yes / no)

   This is how we used “sample” in the last sentence of section 2.1 ( … so the total number of samples is 2502.)

   We agree that this terminology is confusing, since “sample” means two different things throughout the text, and have made an attempt to clarify this possible source of confusion.

We made the following changes to further clarify the two general issues mentioned above:

1. We have included a new figure (Figure 1 in the new document), which:
   * visualizes the start of the measurement (start is at the time of the magnetic excitation)
   * clarifies the discrete data points which lead to the “samples” (now referred to as “data points” or “points” instead of samples in our manuscript, see below)
   * shows how the first data points are discarded because of the measurement artifact (caused by the magnetic stimulation signal itself).
2. We added a description of when the measurement starts to Section 2.1, along with a more extensive explanation of the digitizing rates.
3. In the context of sampling from a time series, we replaced “samples” with “data points” or “points”.
4. We changed the first mention of samples in the machine learning context to input-output pairs, and clarify that we refer to input-output pairs as samples in the rest of the article. Namely, we added in Section 2.2 (line 94): “It furthermore gives us more input-output pairs for training the model, where the input is a collection of measurement variables (e.g. latency, peak-to-peak amplitude, age), and the output is the disability progression target (yes or no). From now on we refer to input-output pairs as samples.”
Reply: specific comments

Reviewer 2

COMMENTS FROM THE REVIEWER (NEW):
I (still) don’t get a clear idea about when the calculation is performed. When authors referred to "1850 samples", what does it mean, in terms of range of response in the MEP? Because, it is probably a huge task to define this interval for each subject, at least it could be defined as a range in the time window where extraction was performed.

ANSWER FROM THE AUTHORS:
As mentioned in the “general issues” section, we suspect the use of the term “samples” to be the main source of miscommunication. The “1850 samples” refers to the 1850 data points one gets when sampling at 19.2 kHz for 100 ms, and discarding the first 70 points. This is just the time series, or the waveform, as shown in the new Figure 1.

To avoid any misunderstanding, we summarize the measurement process here:
The measurement starts the moment the motor cortex is stimulated. The electrical activity is recorded in either the hands or the feet, roughly once every 0.05 milliseconds (a sampling rate of 19.2 kHz) for a duration of 100 milliseconds. This then results in a time series such as the one shown in (the new) Figure 1, which consists of 1920 data points. Specialized nurses then annotate the latency, based on this timeseries (indicated as ‘3’ in Figure 1). The peak-to-peak amplitude is extracted simply by subtracting the highest value from the lowest value in the time series. Since there is usually an artifact at the very start of the measurement we discard the first 70 points of the time series before we move on to feature extraction.

We then calculate a number of time series features from the 1850 remaining data points (the time series), to which we refer as “feature extraction”. These include things such as the average of all the points, the highest value, the standard deviation of the points, the approximate entropy of the time series, … In total we extract some 6000 of such features. These features are calculated on the whole time series (all 1850 points). It’s to these 6000 features that we apply machine learning to determine suitable candidates that may help in the prediction of disability progression.

(As a remark: one of the main contributions of the article is that this is the first time that features extracted from the *whole* time series have been analyzed, as previously only the latency and peak-to-peak values were extracted, while the rest of the time series was thrown away.)

We hope this is now clear from the changes in the text, and the addition of a new figure (Figure 1).

COMMENTS FROM THE REVIEWER (NEW) [CONTINUED]:
Another concern reading the 2.1 section is about that some subjects showed more than one measure, so the calculations were not weighted between subjects. What’s the reason of this procedure?

ANSWER FROM THE AUTHORS:
If we understand correctly, “more than one measure” refers to patients having more than one input-output pair included for training the model. Here, input-output pair means: input = variables to train the model (e.g.: latency, peak-to-peak amplitude, age, …); output = disability progression after 2 years yes / no. Because some patients have a longer follow-up period (e.g. 15 years), there could be several
"disability progression after 2 years yes / no" outputs. On the other hand, a patient that has been in follow-up only for 2 years has exactly 1 input-output pair. By not reweighing between subjects we are, indeed, making subjects with more than one input-output pair more influential. Our assumption is that the machine learning models benefit from including these input-output pairs because:

* they significantly increase the total number of input-output pairs, which is beneficial for training the model
* the total number of input-output pairs for one patient relative to the total number of input-output pairs in the training data is small, so we do not think it is likely they “overinfluence” the model much.

More specifically, if the training data is 80% of the full dataset, the majority of patients represent between 0.05% and 0.5% of the total training set, while there are 2 outliers which represent around 0.9% of the training set.

Note that all input-output pairs of one patient occur either fully in the train or test set. Therefore, the net effect (if any) of our current method would be to lower the performance of the model compared to models where subjects are reweighted, making our performance estimates conservative. Indeed, if the model puts too much emphasis on a few patients, generalisation to new patients will be worse.

REVIEWER COMMENT (NEW):
If the full time series are used, are you also including time before the execution of the response (as a baseline)? And, if authors stated the following: "Because each EPTS starts with a large peak at the beginning, an uninformative artifact of the electro-physiological stimulation, the first 70 samples of each EPTS are discarded." When it is supposed that the extraction starts? The problem to understand all this procedure is that, at least for me, I could not replicate this study because no indications of latencies or peak to peak measurements have been made.

ANSWER FROM THE AUTHORS:
We hope to have clarified these questions with the new Figure 1 (discussed previously in our reply), and in the “Reply: general issues” section.

The measurement starts when the magnetic stimulus is applied. We added a description of when the measurement starts to Section 2.1. The baseline (= the signal between the stimulus in the motor cortex and the response in the hand or feet) is included, and should be flat. This is shown in the new figure 1.

REVIEWER COMMENT (NEW):
The requested information has not been included. I strongly suggest to the authors that take a specific article about ERPs and the usual data included on it (for instance, Coppola G1, Currà A, Di Lorenzo C, Parisi V, Gorini M, Sava SL, Schoenen J, Pierelli F. Abnormal cortical responses to somatosensory stimulation in medication-overuse headache. BMC Neurol. 2010 Dec 30;10:126. doi: 10.1186/1471-2377-10-126).

ANSWER FROM THE AUTHORS:
We hope to have clarified this issue in the “Reply: general issues”.

We have identified an error in section 2.1: The sampling rates of the machines were 20kHz and 19.2kHz respectively, not 50kHz and 48kHz as was claimed before. Perhaps this detail has led (or contributed) to the confusion regarding the samples.
REVIEWER COMMENT (NEW):
The goal of the study is quite interesting, however, in my opinion, the manuscript is still not clear enough in its contributions to the neurology field.

ANSWER FROM THE AUTHORS:
We hope that this issue is partly or completely resolved by our clarifications on the type of measurement (motor EP) and pre-processing in the new version of the manuscript.

To summarize our contributions:
[1] In the previous literature, features studied from the motor EP were limited to the latency, peak-to-peak amplitude, and sometimes a manual rating of “morphological abnormality”. These features were, furthermore, often condensed into one single variable: the EP score. Such a procedure could miss interesting information included in the whole EP time series.
We analyze, for the first time, features extracted from the whole time series, which allows us to look for interesting time-series features (e.g. mean of all data points, kurtosis of all data points, or more complex measures such as the sample entropy) that are contained in the whole time series but have not yet been studied. The inclusion of these features could provide extra feedback during patient follow-up, which is relevant for neurologists treating MS patients. For example, the best feature we found for the measurements in the feet (AH) is marked as important more consistently than any of the other features currently commonly used in the literature (latency, EDSS at T0, age). We’ve added a few sentences to the conclusions to emphasize this result, as well as to the discussion of said feature in Section 3.3.
[2] We study a task that is performed often in the EP literature: predicting disability progression of MS patients (here, after 2 years). We have, for the first time, a dataset that is large enough to get a performance estimate on an independent test set, containing different patients. This performance measure is much closer to clinical practice (where the algorithm should work on new patients). As such, it provides a significant step towards bringing statistical (machine learning) models on EP data to real-world clinical practice, which is relevant for MS neurologists. This is especially true since our performance estimates show that MEP have good predictive performance, meaning that further investigations into their use for clinical follow-up are interesting.

We hope the relevance to the neurology field is clear from the last 3 paragraphs in the Discussion section.

Reviewer 3

COMMENT TO THE ANSWER
The hold-out evaluation (train/test split) is good if you have a very large dataset. Some of the reasons are:
- Cross validation is time-consuming
- If the dataset is large enough, no repetitions of the training/test datasets are necessary.

However if the authors find a high variability I would recommend to use stratified holdout to ensure that the classes are equally represented in the samples.

ANSWER FROM THE AUTHORS:
We agree with the comments of the reviewer that the classes should be equally represented in both the train and the test set. We ensure that patients don’t occur both in the training and the test set, and that
the balance of the targets is roughly the same for the train and the test set. We refer to this as grouped (by patients) stratified shuffle splits in Figure 2. The shuffle pertains to the fact that the samples are distributed randomly across the train and test set for each split, subject to (i) stratification and (ii) patients occurring in only the train or test set.

This is outlined in Section 2.3.

COMMENT TO THE ANSWER
The author claim that a 5-fold cv is unstable. However they use a 4-fold cv on training set to choose the hyperparameters (see Figure 2). It makes no sense to me.
Furthermore, they claim that they achieve with 5-fold cv an accuracy of (i.e.) 0.72 (0.07) and if you see Table 2, the first result is an accuracy of 0.73 (0.08). Results are similar and the variability too. So, why cv is unstable and holdout (train/test) not?

ANSWER FROM THE AUTHORS:
A more detailed answer to the comment is given below, but we already give the conclusion here:
5-fold CV and 1000-split hold-out validation are indeed similar in their variability (standard deviation) and in their estimate for the average AUC. However, we need to consider more than 5 splits to reduce the Standard Error on the Mean performance (a.k.a. the SEM, not to be confused with the standard deviation), which is why we consider 1000 splits. We need this SEM to be low to get a good estimate of the performance increase which occurs when we add the additional time series features.

Detailed answer:
One of the main goals of our work is to show that adding additional time series features increases the performance of a prediction model. We therefore need to quantify this increase in performance. Preferably, this increase should not depend on the way the train/test splits are constructed. This is usually why one employs CV: to reduce the impact of the choice of splits. This gives us 5 performances, one for each of the 5 different ways of constructing the train/test set. The average of these is an estimate of the real average performance across all possible ways of constructing the train/test splits. But since one considers only 5 splits this estimate of the real average is not particularly accurate, and it will still vary considerably depending on which 5 samples are chosen. This variability is quantified by the Standard Error on the Mean (or SEM, see, e.g., Douglas G Altman and J Martin Bland, BMJ. 2005), which is the variability we were talking about in our previous answer. We should have communicated this more clearly: calling 5-fold CV unstable in our previous answer was a confusing choice of words, since both 5-fold CV and hold-out validation have similar standard deviation. The SEM gets smaller as the number of splits increases, following the rule:

\[ \text{SEM} = \frac{\text{sigma}}{\sqrt{n}} \]

where sigma is the standard deviation of the performances, and n is the number of splits. Plugging in some numbers it becomes clear why we had to increase the number of splits under consideration:

\[
\begin{align*}
\text{SEM 5-CV } &= \frac{0.08}{\sqrt{5}} \approx 0.04 \\
\text{SEM 1000-HO} &= \frac{0.08}{\sqrt{1000}} \approx 0.003
\end{align*}
\]

where 1000-HO indicates the 1000-split holdout. If we wish to quantify the performance increase due to adding time series features to the model, we get:

5-CV: Delta AUC = (0.745 +- 0.04) - (0.725 +- 0.04) = 0.020 +- 0.06
1000-HO: Delta AUC = (0.745 ± 0.003) - (0.725 ± 0.003) = 0.020 ± 0.004

where the ‘+–’ indicates the SEM in this case. The result we get from the 5-CV is not as compelling as the one we get using 1000-HO. The high SEM of 5-CV is what we called ‘unstable’ in our previous answer. To illustrate this, in our previous answer we calculated the values obtained from the 5-CV for 3 different ways of constructing the folds. Indeed we see that for the 3 seeds that were considered we get Delta AUC’s of 0.033, 0.0114 and 0.019 respectively.

It should be noted that these are only rough estimates of the SEM as our splits are not truly independent due to overlap of patients in the splits. As a result, the scaling is not exactly 1 / \sqrt{n}.

So we can conclude that while 5-fold CV is enough to determine that adding extra time series features increases performance, it is not enough to get a sufficiently accurate estimate of the absolute value of the increase. For this more splits are required, which is why we use 1000 splits. We’ve added the following sentence to Section 2.3 to make this more clear in the manuscript:

“It also drives down the standard error on the mean performance estimate, allowing for a more accurate quantification of the performance increase we get by adding additional time series features to the model.”

We also clarified the use of the ‘+–’ in the manuscript to mean the standard deviation.

Regarding the 4-fold CV used to determine the values for the hyperparameters: In this scenario there is no need for an accurate estimate of the absolute difference in the performance of the model for 2 values of the hyperparameters, just knowing which value performs better is enough. Therefore, 4-fold CV is sufficient to determine these values.