Review of: "Predicting vertical ground reaction forces from 3D accelerometery using reservoir computers leads to accurate gait event detection"

Thierry Gosseye[1]

1 Université Catholique de Louvain

Potential competing interests: The author(s) declared that no potential competing interests exist.

Dear Madam, Sir,

I have read this manuscript with great attention. This was a pleasure to discover your work in first hand. I'm also convinced that machine learning combined with inertial sensors can bring really interesting information on numerous activities out of the lab environment. We live a period with fast development of sensors in signal quality and device miniaturization. Together with computation power of current computers, the possibilities are infinite. So it is important to choose with care the computation done to bring true progress to science. The introduction of the manuscript could emphasize the importance of measuring FC and FO time to justify the amount of work done. In my opinion, a machine learning process that is specially intended to predict FC and FO, not computing GRF first, should be able to outperform your algorithm. However the computation of the GRF is at least as important as stance timing in the evaluation of locomotion. For this reason, your work is of great interest for the scientific community.

You state that it is a proof of concept, given the low number of subject and conditions tested. You seem highly confident in the generalization of your algorithm. But the precision of your results are probably due to the similarity of the condition tested. I admit that predicting walking and running pattern with a single algorithm show that the process that you have used is quite powerful and robust but the result will be worst with the large panels of situations that can be encountered out of a treadmill: acceleration, slope, turn, BMI range, speed range, uneven ground, training, pathologies. The manuscript could state that a validation with a wider panel of conditions will give more confidence in the results. Measuring the true GRF in more conditions with force-plates is a large work but the FC and FO time errors can be evaluated with simple pressure insoles.

The testing that is done is biased as soon as the data of a test subject are split in the training and in the testing group. The data in these groups should be independent so from separate subjects. If it is not the case, the testing does not reflect the performance on a new subject not used in the training phase. See DOI: 10.1016/j.jbiomech.2018.09.009.

Here are more specific comments along with some questions to ensure that I understood well everything.

Lines 34-36: Feature extraction is a common way to proceed. The double integration that you make is a feature extraction. It does not induce bias if done correctly.

Line 40: As I said above, it is not obvious to me, but it is not a problem.

Lines 46-48: I understand from this sentence that your method manage to predict the GRF during the swing phase while other methods don't. It is obviously zero. I understood after reading the entirety what you mean. It could be clarified.

Line 42: Why did you choose shank? Is it recommended in some reference or is it the result of a thought from your team or because of freely available data.

Lines 62-89: It looks like a method paragraph when the reader starts to read the result chapter. However I understand that you have to explain a bit of the methods before the results since you placed the methods in the end.

It seems logical after reading everything that you predict the left GRF from left shank accelerometer and the right GRF from right shank accelerometer, but arriving at this point of the manuscript it is a question in the mind of the reader. It could also be clarified in this section that the 100 random draw are a random draw of trial and a random draw of the reservoir matrix C and F.

Line 91: I understand that this SD related to the variability of the average error among the 100 draws. Is it true or is it related to the variability of the error in trials? How is normalized this error? By peak value, by average value, … ?

Figure 2: How are placed the dots on the violin plot? Ok for the ordinate but at which abscissa?

Lines 137-147. This paragraph could be part of the introduction.

Lines 148-153: This is an interesting point and can be placed in the introduction as driver for this study.

Lines 165-167: Concerning the pre-processing, I'm wondering if the velocity and displacement obtained by integration are not implicitly included in the output of the reservoir? Does it strongly improve performance to include this pre-processing?

Line 217: Predicting clinical gait require specific validation on the pathologies involved. Clinical data has large variability so are more difficult to predict. Different articles show the lower performance of machine learning on clinical data, for example DOI: 10.1016/j.gaitpost.2019.07.190.

Lines 219-221: Treadmill running mimics quite well overground running while in steady speed on flat ground or on slope, not during acceleration, turn, or soft ground. These conditions are met when looking for ecological environment.

Conclusion: The conclusion should state that it is a pilot study that is not validated on a wide panel of conditions, so the performance should be taken with care.

Line 250: Where the device was placed on the tibia? Is it on a specific muscle to record EMG or somewhere with little soft tissues to minimise soft tissue artefact? This can impact results.

Lines 256-257: These SD are quite low and this is probably in favour of a good prediction result.

Line 267: Why do you normalise by range? This has more noise than SD normalisation that you use in line 272.

Line 276: This is just a moving average if order one is used isn't it? It results in a cut-off frequency of 15Hz with the sampling frequency of 142Hz. That is not a very weak filter and can have an impact on results. Did you optimize the filter? I would have used a Bessel filter that shows less distortion of the signal.

Lines 279-281: The 36 samples represent 252ms. Did you verify that this is enough for node stabilization? The 36 samples after foot off do not seem useful to me. The output GRF is not impacted by future accelerometer data if I understand well the prediction algorithm.

Line 284: The list of q values should finish by q indices N.

Line 296: I'm not a native English speaker but I would say "to-be-learnt" in place of "to-be-learning".

Lines 302-304: Is there guidelines that you followed for the setting of these parameters? Are they optimized manually or in an automatic procedure? Do these parameters have an important effect on the performance? Also, I do not see how adding noise prevent overfitting unless you repeat the same data multiple times with different random noise. Is it what you have done?

Line 322: By using the word “100 iteration”, you suggest that there is an optimization of some parameter from one iteration to the next. I understood that it is 100 repetitions of the process with no relation between each other.

Line 334: In the leave-M-out cross-validation, I understood that the testing group is M trails but you write here that it is 25% of data.