"WARNING: A wearable inertial-based sensor integrated with a 2 support vector machine algorithm for the identification of 3 faults during race walking" by Juri Taborri, Eduardo Palermo, and Stefano Rossi

Summary:

The sport of race walking currently exists with subjective refereeing. Thus, race results are frequently questioned. In at attempt to provide more objective feedback in the sport, this manuscript and experiments attempts to detect faults using IMUs. 10 expert race-walkers were included in the experiment (8 male, 2 female). Three conditions were observed legal walking, illegal – loss of contact, and illegal – knee bent. Three categories of machine learning algorithms: decision tree, SVM, and kNN were evaluated. Evaluations were on accuracy, F1-score, and G-index. Speed of prediction was also considered. Quadratic SVM was the best classifier with an accuracy above 90%.

Comment/consideration for authors:

I do not presume to be an expert in race walking or the implications of this device. However, I wonder, from an outsider's perspective, if the ultimate goal of this device should be to minimize cases where a fault is detected but no fault actually occurred (rather than focusing on the more traditional measures of general effectiveness like F1-score, G-score, etc). This seems plausible to me because in application, this device would, presumably, be used to authenticate race results, including potentially disqualifying participants that accrue some number of faults. In this scenario, a false detection would be devastating for a racer that may otherwise have won. Missing a single fault, however, seems unlikely to be deterministic in the results of the race. Racers simply knowing that faults may be autonomously detected should deter intentional faulting, and participants who regularly fault would still have some faults detected. With an accuracy of roughly 92%, that means that 8 out of every 100 strides taken would be misclassified, and with an F1 score of .92, it seems as though this would be roughly equally likely to misidentify a fault as a non-fault and a non-fault as a fault. So, in any sort of extended race, the number of faults detected would essentially still be meaningless, despite this high accuracy because the number of strides taken is so large. This would, practically speaking, make this device unusable. However, with a device that never incorrectly detects a fault but does correctly detect some faults, there may be a practical application for this.

Review:

The work does overlap some with the authors' previously published work. However, this work includes two additional participants in the dataset (or perhaps 10 new ones? This is not clear.) and evaluates a more generalized (labeled as "inter-athlete") solution. It also evaluates a wider range of approaches.

Some claims made throughout seem misleading, incorrect, or overstated. Two examples below, though there are other, similar statements in several locations:

1) Page 8, lines 286-287: “Thus, it is assured to avoid misclassification, both in terms of false positive and false negative.” – this is false. Misclassifications are certainly still possible. The only way to avoid misclassification at all would be with a 100% accuracy, which is not the case here.
2) Page 9, line 351-352: “… allow reaching optimum values for all the three parameters; it assesses the possibility to use only one sensor accepting a limited reduction of the performance.” – depending on interpretation, this could be true, but what “optimum” means can cause this to be misconstrued. (note that this single sentence has several grammatical issues. One example: “all the three parameters” should just be “all three parameters” grammatically). This sentence should be rewritten to more clearly indicate that there was a performance decrease when one sensor was used and that statistically speaking, there is not a significant difference between which leg was instrumented with the sensor. This requires support from a p-value being above 0.05 when comparing R and L directly. If the p-value shows a statistically significant difference (it is < 0.05), then that leads to a different discussion about why this occurred.

It is stated (page 4, line 185 and in 4 other locations) that there are three datasets. Unless I am mistaken, there is 1 dataset. The dataset including both legs is the full dataset, and the study also examined how the classifiers would perform with a limited version of the dataset using only one sensor (rather than both) as well. I believe this discrepancy amounts to a terminology issue, but one that could be misleading to readers. 3 datasets implies three separate experiments or sources of data rather than one source with two different ways of subsampling the data.
Discussions of limitations in the study and future work are not discussed in enough detail.
There are many English language grammatical issues and spelling issues. All sections should be reviewed carefully for grammatical issues. Example(s) of each:
Spelling:
page 7, line 268: significative -> significant
page 9, line 349: hystogram -> histogram
page 9, line 350: otpimize -> optimize
Grammatical:
page 3, line 125: “we used an offline processing” -> “we used offline processing”
page 8, lines 283-284: “… it is evident as almost the totality of the tested classifiers can be considered as an optimum classifier, …”