Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion

Tim Op De Beéck
Dept. of Computer Science,
KU Leuven
Leuven, Belgium
tim.opdebeeck@cs.kuleuven.be

Wannes Meert
Dept. of Computer Science,
KU Leuven
Leuven, Belgium
wannes.meert@cs.kuleuven.be

Kurt Schütte
Human Movement Biomechanics
Research Group, Department of
Movement Sciences, KU Leuven
Leuven, Belgium
kurt.schutte@kuleuven.be

Benedicte Vanwanseele
Human Movement Biomechanics
Research Group, Department of
Movement Sciences, KU Leuven
Leuven, Belgium
benedicte.vanwanseele@kuleuven.be

Jesse Davis
Dept. of Computer Science,
KU Leuven
Leuven, Belgium
jesse.davis@cs.kuleuven.be

ABSTRACT

Running is extremely popular and around 10.6 million people run regularly in the United States alone. Unfortunately, estimates indicated that between 29% to 79% of runners sustain an overuse injury every year. One contributing factor to such injuries is excessive fatigue, which can result in alterations in how someone runs that increase the risk for an overuse injury. Thus being able to detect during a running session when excessive fatigue sets in, and hence when these alterations are prone to arise, could be of great practical importance. In this paper, we explore whether we can use machine learning to predict the rating of perceived exertion (RPE), a validated subjective measure of fatigue, from inertial sensor data of individuals running outdoors. We describe how both the subjective target label and the realistic outdoor running environment introduce several interesting data science challenges. We collected a longitudinal dataset of runners, and demonstrate that machine learning can be used to learn accurate models for predicting RPE.

KEYWORDS

Machine Learning, Sensor Fusion, Sports Analytics

ACM Reference Format:
Tim Op De Beéck, Wannes Meert, Kurt Schütte, Benedicte Vanwanseele, and Jesse Davis. 2018. Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219864

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219864

1 INTRODUCTION

Worldwide, recreational running is one of the most popular forms of physical activity. In the United States alone, 10.5 million people run regularly, and around 36 million people in total participate in running each year [20]. While running regularly has many health benefits, injuries hamper these benefits and can even be detrimental for the runner. Unfortunately, runners are prone to overuse injuries, with estimates indicating that anywhere from 29% to 79% of runners suffer at least one overuse injury per year [31]. Overuse injuries arise due to the repetitive nature of the movements performed during running. These movements repeatedly stress (i.e., apply force to) the same structures (e.g., muscle tissues, tendons or joints) in the body. The effects of the stress accumulate over time, and may eventually exceed the structure's stress tolerance, resulting in an injury [12, 18]. Typical overuse injuries in running include pain under the foot (plantar fasciitis) and pain on either the front (patellofemoral pain syndrome) or side (iliotibial band friction syndrome) of the knee [29].

While current research is inconclusive, three categories of factors have been linked to overuse injuries. First are anatomical factors such as high arches, which are inherent to a person. Second are training factors such as excessive long-distance running. Third are biomechanical factors such as the symmetry between the right and left side in a person's running movement. For the third factor, the onset of running fatigue plays a crucial role as it can alter a person's running style, that is, the movement pattern performed from one step to the next while running. Changes in style can introduce irregularities such as asymmetries between the left and right side, which can elevate the risk of an overuse injury [26]. Because these movement alterations are very subtle, they are often not consciously observed by the runner.

Thus, *predicting an individual's fatigue state* based on his currently observed running style has the potential to reduce the risk of overuse injury. While this task can naturally be posed as a supervised machine learning problem, several factors make this an extremely challenging task. First, we need to monitor and characterize running style in "the wild", that is, in a real-world outdoor setting (e.g., variations in weather, running speeds, etc.) in contrast

to traditional, controlled laboratory conditions (e.g., running at a fixed-speed on a treadmill). Second, due to several inherent physiological and morphological differences, individuals will respond differently to the same type of exercise. Third, measuring the fatigue state is highly non-trivial. Within the sport science literature, researchers distinguish between different types of fatigue, such as cardiovascular fatigue, biomechanical fatigue, respiratory fatigue, or mental fatigue among others. Which type of fatigue is relevant depends on the task. While heart rate can capture cardiovascular fatigue, overuse injuries are related to biomechanical fatigue. Therefore, we focus on measuring biomechanical fatigue, which can be invasive and expensive (e.g., blood lactate), or represent a subjective measurement of fatigue (e.g., rating of perceived exertion or RPE).

In this paper, we present a machine learning approach for predicting a runner's RPE, a subjective fatigue measure, based on fusing inertial motion data. We introduce this as an interesting and important data science challenge. In particular, it involves challenges such as analyzing noisy real-world data, handling partial ground truth labels, and reasoning about subjective judgments that vary over time. Our approach is based on defining a variety of biomechanically relevant features that characterize a person's running style. We then build regression models to predict the RPE value at a specific point in time. We evaluated our approach on a longitudinal data set of 29 runners. Each subject completed at least three maximal effort running tests on an outdoor track while wearing four inertial motion units (IMU). We found that, on average, we are able to accurately predict a runner's fatigue state. We found no substantial benefits to fusing the data from multiple sensors compared to using inertial motion data captured from the wrist. Furthermore, we showed that we could effectively deal with the subjectivity of the target variable and the noise introduced by variable running speeds and inter and intra individual differences.

To summarize, this paper's contributions are:

(1) Introducing fatigue prediction in runners as an interesting and important data science problem;
(2) Highlighting a number of data science challenges that we encountered while working on this problem;
(3) Describing a supervised learning pipeline for this problem that addresses these challenges;
(4) Presenting the results of predicting RPE on a real-world longitudinal data set; and
(5) Illustrating that several techniques can, to some extent, account for the subjectivity of the target variable and inter and intra individual differences.

2 FATIGUE PREDICTION: DEFINITION AND DATA SCIENCE CHALLENGES

This paper aims to solve the following problem:

Given: Multiple signals collected by inertial sensors placed on a runner.
Do: Learn a model to predict the runner's fatigue state at a given point in time.

This section begins by defining fatigue and how to measure it, then describes our data, and finishes with a discussion of the challenges posed by this task.

2.1 Measuring Fatigue

The first issue is measuring an individual's running-based fatigue state, which can be thought of as a hidden variable with a continuum of possible values. Running induced fatigue implies a decrease in running performance (i.e., decreased average speed) due to physiological limitations (i.e., low aerobic capacity, low lactate threshold, or poor running economy) that bring about biomechanical compensations (i.e., alterations in the running kinematics). Several possibilities or markers exist for capturing a runner's fatigue state, yet not all of them are appropriate or suitable for our task. Hence, we developed four primary criteria for selecting the most ideal measure of running fatigue:

Non-invasive. That is, the measurement method or device does not involve the introduction of instruments into the runner's body. Examples of invasive measurements of fatigue include blood lactate [28], creatine kinase [15], or rectal temperature [6].

Unobtrusive. That is, the measurement method or device does not hinder the runner's comfort in any way and does not interfere with the fluidity of the runner's movement. For instance, obtrusive measurements may include metabolic systems that measure gas exchange (i.e., volume of oxygen consumed (VO_2) or carbon dioxide produced (VCO_2)). Although some of these more portable metabolic systems are wearable, they require a constrained harness, a heavy battery pack, and an uncomfortable face mask that often hinder a runner's comfort.

Non-interruptive. That is, collecting the measure does not interfere with the runner's performance or continuity. Interruptive measures would include both invasive such as blood lactate, as well as non-invasive measurements such as heart rate variability which is known to be inaccurate during dynamic activity [7]. Interruptive also implies unnecessary physical or mental effort is required by the runner. For instance, more sophisticated rating scales that subjectively quantify fatigue include the *Hooper's Index* [11] or the *profile of mood states (POMS)* [33], which are time consuming and require cognitive loads that force measurements to be attained prior or post running.

Fatigue Specificity. That is, while running the measurement or device provides insights into the musculoskeletal response, which has closer links to overuse injury. For instance, at low to medium aerobic intensities, a runner's biomechanical loading can gradually accumulate and movement compensations may arise while HR can remain relatively stable, suggesting a "mismatch" in fatigue between the musculoskeletal and cardiovascular systems. Thus, although other measures such as heart rate (HR) may fulfill the criteria of being non-invasive, unobtrusive, and non-interruptive, it lacks specificity by only providing insights into the cardiovascular, rather than the musculoskeletal response to running.

Consequently, we measure fatigue using the rating of perceived exertion (RPE), a subjective measure of fatigue that is widely used in running research specifically, and within sport science more generally. Specifically, we use the Borg scale [3], where subjects indicate their perception of exertion between 6 (i.e., no exertion)

and 20 (maximal exertion). Because it is subjective, RPE should be viewed a partial truth label. However, RPE has several advantages because it is non-invasive, unobtrusive, and non-interruptive due to its measurement simplicity. Importantly, RPE also has fatigue specificity, given that it provides a more holistic view of fatigue that is said to represent feedback from cardiovascular, respiratory and musculoskeletal systems [6]. Furthermore, RPE has been shown to model a runner's performance better in the real-world compared to heart rate which is less responsive to different terrain types [2]. Thus, RPE is an appropriate and validated marker of a runner's fatigue [3].

2.2 Data

The data used to train our model consists of longitudinal data for 29 runners. In total, data from 98 trials was collected, where 20 runners completed three trials, seven completed four trials, and two completed five trials. Each trial consists of completing a 3200 meter run on an outdoor track (one lap is 400 meters). Each runner was instructed to use a self-selected pacing strategy to run the trial such that they were fatigued by the end of the run and would reach a RPE between 16 and 20 (very exerted). Running outdoors means the test more naturally mimics the running style of a runner's regular training sessions compared to running on a treadmill at a controlled speed. The study protocol was designed in collaboration with biomechanics researchers with extensive expertise in collecting and analyzing running data. From a sports science perspective, this is a larger data set than usual because data collection is very time consuming, with this collection effort taking > 4 months. The study was conducted according to the requirements of the Declaration of Helsinki and was approved by the KU Leuven ethics committee (file number: s59353).

Prior to starting the run, we explained the RPE scale. The scale ranges from 6 through 20 inclusive and the runner is free to pick any integer value in this range. To help the runners understand the scale, we provided verbal fatigue anchors (e.g., 15 = "hard", 17 = "very hard", 19 = "extremely hard") for every odd number on the scale. Then each participant ran one warm up lap followed by the 3200 meter test. The runners reported their RPE after each lap, including the warm up lap, yielding nine RPE values per trial. Figure 1 visually illustrates the protocol. Per lap RPE was thought to be a reasonable time-frame to capture fatigue changes without hampering the runner's performance (e.g., additional mental fatigue or distractions caused by frequent RPE measurements).

During the test, six 1024Hz inertial motion unit (IMU) (Shimmer 3, Shimmer, Dublin) and a strap-based heart rate monitor (Garmin Forerunner 210, Garmin, Schaffhausen) at 1 Hz were attached to the runner. Each IMU contains an accelerometer, gyroscope and magnetometer that measures one signal for each of the three orthogonal axes per sensor type, resulting in nine signals per IMU. One IMU was attached to each of the left and right: shin bone (anteromedial aspect for the distal tibia), wrist (dorsal carpal ligament) and arm (at the level of the mid-point between the acromiale and the radiale, on the mid-line of the lateral surface of the arm). Unfortunately, sensor errors sometimes caused the data from one of the wrist or arm sensors to be lost. Therefore, only data from one wrist and one upper arm sensor was used. If available, we used the left wrist and left arm sensors. Otherwise, we used the right wrist or right arm sensors. Thus four sensors were considered in total. Finally, each lap time was recorded by a hand-held stop watch.

2.3 Challenges in Fatigue Prediction

Using a subjective measure of fatigue as the target label introduces several challenges:

- **C1: Subjectiveness of Target Label.** Different runners will rate their exertion level differently. Moreover, it is often hard for runners to accurately assess the gradual and subtle increases of their exertion level throughout the test.
- **C2: Accommodation to the Test Protocol.** Runners were instructed to run in a way that resulted in a RPE score between 16 and 20 by the end of the trial. Consequently, some runners likely reported a high RPE score at the end because this was the expected behavior, and not a score that reflected their true RPE.
- **C3: Evolution in Reporting RPE.** Most subjects were unfamiliar with the RPE prior to the study, and were perhaps unsure how to use it at first. The longitudinal nature of the study means that runners became more familiar with the scale as they ran more tests, and therefore their use of the RPE potentially evolved across consecutive running tests. This issue is similar to problems associated with working on rating data (e.g., for movie prediction) in machine learning [16].

Employing a study protocol that mimics normal running (e.g., outdoors, self-select speed), introduces a number of challenges into the data that should be accounted for:

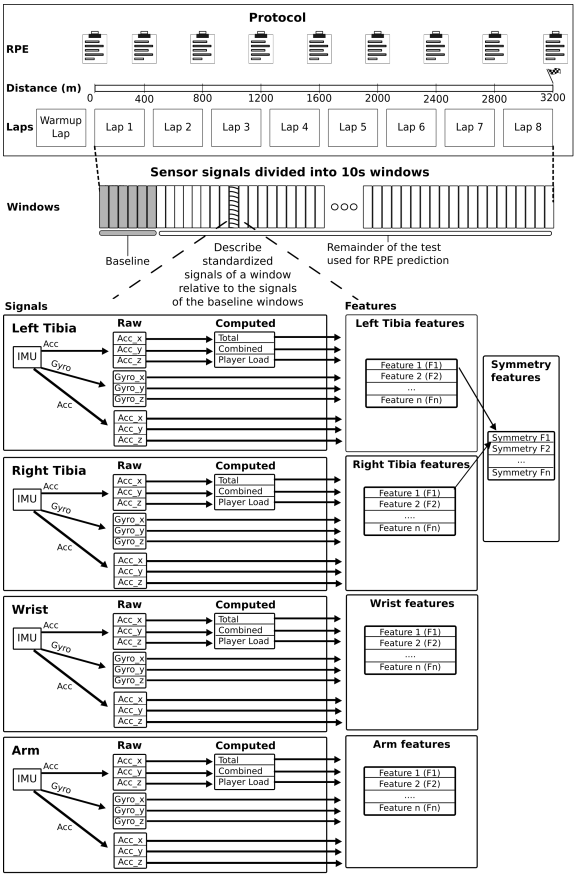
- **C4: Pacing Strategies** Runners apply different pacing strategies during the test. Experienced runners are able to maintain a nearly constant speed over a test. In constrast, many novice runners start fast, slow down in the middle, and speed up at the end. Furthermore, subjects use past experience to alter their running strategy for subsequent tests. Thus there are both inter and intra subject differences in pacing strategies.
- **C5: Variable Running Speed** Running speed, and changes in it, impact the measured inertial motion data (e.g., higher speed means higher acceleration measurements). As we are only interested in fatigue induced changes in the data we need methods that are robust to speed changes.
- **C6: Individual Running Style** Individual characteristics (e.g., weight, height, fitness level, strength, flexibility and training background) mean that runners will have different running styles. These unique styles affect parameters such as step length, step frequency, and arm movement.

3 OUR APPROACH TO FATIGUE PREDICTION

In this section, we outline our approach to fatigue prediction. First, we discuss which signals we consider. Second, we describe how to construct examples and how to address the challenges described in Subsection 2.3. Third, we describe which features we compute for each example. Fourth, we discuss how to build models. Figure 1 provides an overview of our approach.

3.1 Signals Considered

Figure 1: Overview of protocol, data preprocessing and feature extraction



Our data originally contained heart rate, accelerometer, gyroscope, and magnetometer signals. We altered this in three ways. First, as discused in Subsection 2.1, the heart rate was of limited value because it plateaus quickly. Therefore, we omitted the heart rate data. Second, we also omitted the magnetometer data as this signal, in isolation, does not provide information about running style. Third, we augmented the data by deriving five additional signals from the accelerometer data from an IMU sensor that are commonly used in sport science:

Total Acceleration. This signal is less dependent on the exact attachment of the sensor as it combines the x, y and z acceleration signals at time t_i, and is defined as:

$$\sqrt{{a_{x_i}}^2 + {a_{y_i}}^2 + {a_{z_i}}^2}.$$

Combined Acceleration. The following three signals were found to work well for gait identification because they are less sensitive to the device's attachment [9]. Each signal computes the alignment of the accelerations along one particular axis with respect to the total acceleration. We compute these combined signals, by comparing each of the x, y, or z axes to the total acceleration:

$$C(v_i) = \arcsin\left(\frac{a_{v_i}}{\sqrt{{a_{x_i}}^2 + {a_{y_i}}^2 + {a_{z_i}}^2}}\right).$$

Player Load™. This signal was developed by Catapult to monitor the changes of accelerations in team sports using an IMU-unit attached to the upper back [4]. However, to our knowledge, this signal has neither been used for runners nor calculated based on data collected on the tibia, wrist or arm. In contrast to team sports, where player loads are often aggregated over a training session, we compute an instantaneous change at time t_i as: $\sqrt{\frac{(a_{x_i}-a_{x_{i-1}})^2+(a_{y_i}-a_{y_{i-1}})^2+(a_{z_i}-a_{z_{i-1}})^2}{100}}$.

Alternatively, the raw accelerometer signals can be rotated from a device reference system to an earth reference system [19]. As a result of this process, the z-axis is always perpendicular to the earth's surface. However, early experiments showed that this rotation was not valuable for solving this task. Therefore, it was omitted from all experiments. The reason it was not necessary is that in our protocol, each sensor was firmly attached to a limb. Hence, the sensor moved minimally, if at all, during the course of the trial, meaning that the sensor's relative position was fixed. The relative movement contains all the relevant information and it directly expresses movement in, for example, the left-right direction while the rotation would obfuscate this information.

In summary, each IMU generated three raw and five computed accelerometer signals and three raw gyroscope signals. Thus, with four IMUs, each trial is described by 44 time series signals.

3.2 Example Construction and Data Preprocessing

We construct examples by dividing the collected sensor signals into non-overlapping 10 second windows. A 10 second window was chosen because it is sufficiently small to process the data quickly, and represents the typical amount of time used previously with respect to fatigue and running biomechanics [22].

Because runners only report the RPE every 400 meters (mean and standard deviation of lap time: 110s ± 18s and range: 72s–177s), we assign an RPE to each example by linearly interpolating between the RPE values reported at the end of the previous and current lap. RPE is known to linearly change with exercise intensity and running fatigue [3].

To deal with the challenges **C1-C3** related to the subjective nature of the RPE outlined in Subsection 2.3, we apply min-max normalization to the RPE value based on the current test. This normalization helps to account for inter and intra subject differences in RPE. First, each runner may interpret the scale differently and report a different range of values. Second, the first RPE value reported in a trial may serve as an anchor for subsequent ratings in that trial. The minimum value is the RPE reported after the warm up lap (mean and standard deviation of the RPE after first lap: 10.57 ± 1.97 and range; 7-15) and the maximum value is 20, which is the highest RPE value on the Borg scale. We used this value because using

the final RPE value from the test would cause the current label to depend on future data, which is not methodologically sound.

Two other challenges mentioned in Subsection 2.3 are that runners employed different pacing strategies (**C4**) and varied their running speed during the trials (**C5**). To help mitigate the effect of these issues, we standardized every signal within an example. For each signal in an example, we subtract the example's mean value for that signal and divide it by the signal's standard deviation in that example.

On average, each trial generates 78 examples, which results in 7,607 examples in total across all runners and trials. For ten second windows and a sampling rate of 1024Hz, an example consists of 44 time series signals, with 10,240 measurements per signal, and 450,560 measurements in total.

3.3 Feature Construction

We want to define features that describe fatigue-related changes in a runner's style. Specifically, because running is a cyclical repetitive movement, and deviations from a runner's pattern may arise due to excessive fatigue, we want to design features that capture changes in the movement pattern.

We consider three broad categories of features: (1) Simple statistical features, which describe aspects of a runner's movement pattern, (2) more advanced sport science features [14, 21, 30], which capture to what extent a runner is able to copy his movement from one stride (i.e., cyclical motion of one leg) to another, and (3) expert-defined symmetry features [26], which explicitly compare the movement of the left and right leg. While the first two categories are computed by analyzing one sensor's signal, the symmetry features are computed based on two signals (i.e., one from each tibia sensor).

Statistical Features. First, we compute for each signal a set of 15 basic features. We consider four standard features of the signal: *The minimum*, *maximum*, *skew*, and *kurtosis*. We also compute the *average absolute difference (AAD)*, which computes the average absolute difference between each value in a signal and the signal's mean value [17]

We compute ten features based on constructing a *binned distribution* of the signal [17]. The signal is divided into ten equal sized bins based on its minimum and maximum value. There is one feature per bin which is equal to the proportion of the signal's values that fall in that bin.

Additionally, for the total acceleration, we construct two features based on the *time between peaks*, which was found to be a useful feature in activity recognition [17]. Since we only have one activity which is cyclical, a window can be more accurately partitioned into consecutive strides by applying peak detection on the total acceleration signal. The average stride duration and the consistency of the stride durations within a window are then captured by two features: the mean and standard deviation of the stride durations.

Sport Science Features. Second, we compute for each signal three more advanced self-similarity features:

Sample Entropy. This feature measures the complexity of a time-series $T = t_1, t_2, ..., t_n$ as $-log\frac{A}{B}$. Given a length m subsequence in T $seq(x) = t_x \dots t_{x+m}$, B is the number of length m pairs such that $d(seq(i), seq(j)) < r$ where d is the Chebychev distance, and r is a tolerance threshold. Given the set of similar length m pairs, A is the number of pairs that after being extended to length $m + 1$, remain similar (i.e., $d(seq(i), seq(j)) < r$) [24]. Furthermore, it was shown to capture running-fatigue related decline in physiological variability of movement patterns [25].

Detrended Fluctuation Analysis (DFA). This feature divides the signal into segments of equal length l and quantifies the fluctuations of the signal after subtracting local trends (i.e., by fitting a polynomial curve) for each segment. This process is repeated for multiple values of l to plot the signal's fluctuations as a function of l. The feature's value is the slope of the linear curve fitted through these points [5].

Stride Regularity. This feature captures the similarity of consecutive strides. It calculates the value of the first peak in the unbiased autocorrelation signal, which corresponds to comparing the original signal with a copy that was shifted by one stride [21]. The unbiased autocorrelation signal is constructed by varying $m = 1 \dots N$, and for each m computing:

$$\frac{1}{N - |m|} \cdot \sum_{i=0}^{N-|m|} x_i \cdot x_{i+m},$$

where N is the number of data points.

Symmetry Features. Third, we compute symmetry features by fusing the signals from the two tibia sensors to capture to what extent a runner is able to replicate his movement from one leg to the other as asymmetries between the left and right side can elevate the risk of an overuse injury [26]. Specifically, each symmetry feature is computed as the log difference of the absolute value of the single leg feature calculated on the right side and the absolute value of the single leg feature calculated on the left side [32].

Normalization. We express the value of each feature relative to a trial-specific baseline for two reasons. First, we expect gradual changes over time of the feature values relative to a non-fatigued state to capture alterations in running style due to fatigue. Second, individual characteristics may affect the observed signal and hence the derived features. After exponentially smoothing all feature values ($\alpha = 0.4$), we use the first six windows to derive a range (i.e., min and max) for each feature. This represents a feature's baseline value for the runner's starting fatigue state. Using this range, we apply min-max normalization to all subsequent values of the feature. To account for inter and intra individual differences we take the absolute value of the normalized feature values.

Summary. To summarize, each IMU has 11 signals. We compute 15 basic features and three sport science features per signal. For the total acceleration signal, two additional features are derived. This means that there are 200 features per IMU. If both tibia sensors are used, then there are 200 additional symmetry features.

3.4 Learning Models

We consider three different learning settings, each learned based on different subsets of the data:

(1) **All Runners Model (AM).** This setting learns a model using data from all runners. This model attempts to leverage

all the data with the assumption that multiple subjects will have similar changes in style as a response to fatigue.

(2) **Other Runners Only Model (OM).** This setting builds one model for each runner using only data from other runners. That is, no data is used about the runner for whom predictions will be made. The goal of this setting is to assess how accurate predictions will be if we have no training data available for a specific runner. This is interesting because for first time runners, there will not be data. Furthermore, some runners may not provide RPE value, which are needed to train an individual (or group) model.

(3) **Individual Model (IM).** This setting builds one personalized model for each subject using only data from that subject. This model would work well if each subject has a unique alteration in style in response to increasing fatigue.

4 EXPERIMENTAL EVALUATION

The goal of the empirical evaluation is to assess the viability of predicting RPE in a real-world outdoor setting, provide insights into the input data, and discuss the practical impact of the results. Specifically, we address the following questions:

Q1: How accurately can we predict a runner's RPE based on inertial motion signals?
Q2: How does the location of the sensor's placement on the body affect predictive performance?
Q3: Does fusing the data from multiple sensor locations improve predictive performance?
Q4: Can runners rate their RPE consistently and according to the BORG scale?
Q5: Can we further improve the results using more advanced sport science features and expert knowledge?
Q6: What preprocessing steps are important for accurately predicting RPE?

4.1 Experimental Details

We now describe the details of our experiments.

Learners. We evaluated four regression techniques: Gradient Boosted Regression Trees (GBRT), Artificial Neural Network (ANN), Linear Regression with Elastic Net regularization (EN), and Linear Regression with Least Absolute Shrinkage and Selection Operator regularization (LASSO). For all models, we used the implementation available in scikit-learn [23]. For the GBRT, we used the default settings for all parameters except for the following two, because changing them was shown to reduce overfitting [8, 10]: *subsample* = 0.4, *max_features* = 0.9. For ANN, we used the default parameter settings. For EN and LASSO we used the default parameter settings except for one parameter. For EN we tuned the $L1-ratio$ parameter, while for LASSO we tuned the alpha parameter. Both were tuned using five fold cross validation on the training set.

Additionally, we consider two baseline predictors. The first baseline model (MIDDLE) always predicts 13, which is the value in the middle of the Borg scale. The second is a personalized, trial-dependent baseline (TD-Baseline) that always predicts the average of the runner's RPE score after the warm up lap and the maximum value of the Borg Scale (i.e., 20). Both of these models can be thought of as predicting the average of a range (i.e., full Borg scale or trial-dependent) assuming that each value in the range is reported the same number of times. We have to use the maximum value of the Borg scale for the top end of the range because when the trial starts, we do not know what the subject's highest reported RPE value will be for that trial.

All four regression techniques and the two baseline models were considered for addressing Q1 and Q2. For the subsequent experiments, we only considered GBRTs because the results from Q1 clearly indicated that GBRT outperform the other techniques on this task.

Features and RPE. To answer Q1, Q2, Q3, Q4 and Q6, we train all models on the set of statistical features. To answer Q5, we learn models for different combinations of the statistical, sport science and symmetry features. Additionally, we always train the models using the normalized RPE values, except for in Q4 where we consider the original RPE values as well.

Evaluation. We evaluate these models using a cross validation scheme that leaves the last trial of one runner out (i.e., the test set consists of all examples generated from the last trial for one runner). Because our data is longitudinal, this scheme avoids information leakage between the training set and the testing set arising from the future data of a runner appearing in the training set. Moreover, in the Individual Model setting, this ensures that at least two trials can be used to construct the model. All preprocessing (e.g., standardization of the feature values) is solely done on the training data. The predictions of every model are exponentially smoothed ($\alpha = 0.6$).

To assess the model's accuracy, we report the mean absolute error (MAE). Because we train on the normalized RPE values, we need to convert the predicted RPE value, $rpe_{predicted}$, back to the original BORG scale using: $(20 - rpe_{warmup}) \times rpe_{predicted} + rpe_{warmup}$ where rpe_{warmup} is the RPE reported by the runner after the warmup lap. When computing the MAE, there are several factors that may influence the computation. First, the time a runner needs to complete the protocol can vary across trials. Second, within one trial, variations in speed mean that each lap takes a different amount of time to run. Third, runners have completed a different number of trials. As we do not want our calculation to be unduly influenced by one lap, one trial, or one runner, we calculate a global distance based MAE in two steps. First, we compute for each running test the MAE per lap by assigning each window to a lap. When a window spans two laps, we assign it to the lap in which the majority of the time resides. Second, for each runner, we then compute the average MAE over all laps of that runner. The global MAE is then calculated as the average MAE over all runners. The first step, accounts for different pacing strategies and for variable running speeds within a test, that result in variable lap durations. The second step, accounts for the fact that some runners completed more than three tests.

4.2 Experiment and Results for Q1 and Q2

The purpose of this experiment is two-fold. First, we want to evaluate the predictive performance of each learner and each learning setting, that is, the All Runners Model (AM), Other Runners Only Model (OM), and Individual Model (IM) on this task. Second, we want to evaluate the efficacy of the different sensor locations (i.e.,

the arm (A), wrist (W), and tibia (T)) as it is unclear from the sport science literature where sensors should be placed.

Table 1 shows the MAE for all learned models. In terms of learners, GBRTs consistently outperform the other approaches irrespective of the model or sensor location. ANNs clearly perform worse on this task compared to the other learners, probably because ANNs typically require very large amounts of training data (i.e,. more than the 7000 examples we have). The higher MAEs of the LASSO and EN models compared to the GBRT models suggest that the movement alterations are better captured using non-linear relationships between the features and the target or that the GBRT technique is more effectively dealing with the high number of features. It is also reassuring to see that the GBRT model outperforms both baseline models in all nine scenarios. This illustrates that the performance of these models is non-trivial.

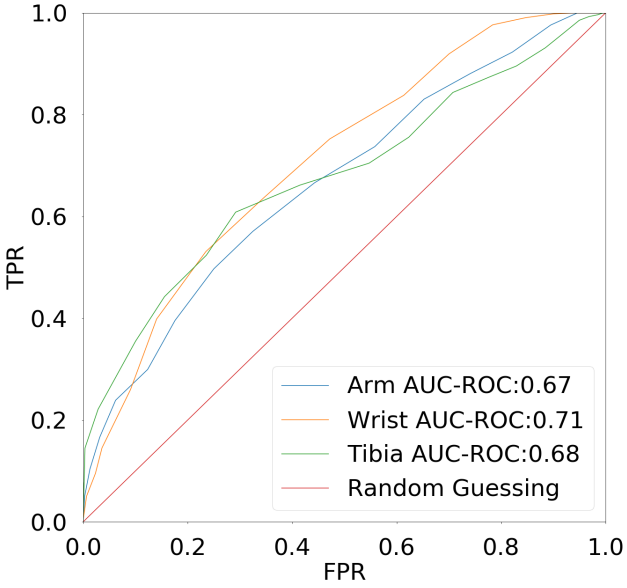
From a more theoretical standpoint, reaching an MAE of 0 is probably not realistic either, as the RPE is a holistic measure that simultaneously captures cardiovascular, respiratory and musculoskeletal fatigue, whereas the IMU sensors only measure musculoskeletal movement patterns.

From a learning setting perspective, using data from all runners results in the best predictions, with slight decreases when only data from other runners is considered. In the vast majority of cases, learning an individual model results in worse predictive performance. The predictive performance of the learners in the AM and OM settings are encouraging as it may be difficult to collect large amounts of data for any given individual in practice. Thus, even if a runner provides no labeled fatigue data it is possible to make reasonably accurate predictions. Furthermore, our methodology already takes into account constraints required by the future real-time prediction system (i.e. no future data is used and many of the calculations are parallelizable).

In terms of sensor location, GBRT achieve the best performance using data from the wrist and has slightly worse results on data from the tibia and arm. This contrasts to LASSO and EN which do better on the arm and tibia than the wrist. Practically, it is encouraging that data from the arm and wrist results in accurate predictions as these are locations where a runner may commonly wear a sensor, either in the form of a watch or an attachment of the smartphone to the arm. In contrast, attaching a sensor to the tibia is less common outside of lab setups and possibly more cumbersome, as it might cause pain to the shins and runners may bump into the sensor with their opposite foot while running.

To evaluate the learned GBRT models in a classification setting, we constructed ROC curves by thresholding the predicted RPE to make a fatigued versus not-fatigued prediction. For the ground truth fatigue state, we considered any reported RPE greater than or equal to 16 as representing a truly fatigued runner. This rating corresponds to hard to very hard on the Borg scale. Figure 2 shows the ROC curves for the GBRT models learned in the AM setting for each of the three sensor locations. The computed AUC-ROC scores show that all three sensors perform similarly for classifying between *non-fatigued* and *fatigued*. Because each point on a ROC-curve corresponds to a threshold for distinguishing between *non-fatigued*/*fatigued*, the selected threshold could be set according to the individual needs of the runner. When selecting a threshold in practice, the following are some important considerations. From an injury prevention perspective, a runner might be better off with a threshold that corresponds to a higher TPR at the cost of a slightly higher FPR. As a consequence, the runner will sometimes be advised to stop running before it is actually necessary. Such a threshold choice would be particularly advisable for novice runners, as most running injuries are mainly due to running too far, too fast, too soon [1]. More advanced and competitive runners could be less conservative and use a threshold that results in a lower FPR at the cost of a slightly lower TPR.

Figure 2: ROC Curves for classifying a runner as being either not-fatigued or fatigued. The results are for the All Runners Model trained using only the statistical features and GBRT for the Arm, Wrist and Tibia sensor locations.



4.3 Experiment and Results for Q3

We hypothesized that movement alterations affect the movements of the legs, wrists and upper arms simultaneously while running. Therefore, we assume that training the GBRT models using features constructed from multiple sensor locations will improve predictive performance. To test this hypothesis, we learned one GBRT model for each learning setting and for each combination of sensor locations.

Table 2 shows the MAEs of these models. Combining multiple sensors seems to result in slight improvements in predictive performance, in each learning setting. Practically, there is a trade-off between improved accuracy and the convenience and cost of wearing multiple sensors. It is unlikely that many recreational runners will buy and wear multiple sensors during each run. Therefore, it is reassuring that there are no substantial benefits to fusing the data from multiple sensors. In the future, the evolution of e-textiles

Table 1: The MAE for predicting RPE for all possible combinations of the four learners, three sensor locations and three learning setting. The three learning settings are the All Runners Model (AM), Other Runners Only Model (OM), and Individual Model (IM).

		AM	OM	IM
Model	**Sensor**	**MAE**	**MAE**	**MAE**
GBRT	**Arm**	1.99	2.03	1.98
	Wrist	**1.89**	2.04	2.15
	Tibia	1.98	2.08	2.02
ANN	**Arm**	2.92	3.32	14.16
	Wrist	6.48	5.54	19.04
	Tibia	4.37	4.5	42.93
ELASTIC NET	**Arm**	2.28	2.34	2.90
	Wrist	3.16	3.24	2.38
	Tibia	2.09	2.11	3.66
LASSO	**Arm**	2.33	2.38	2.94
	Wrist	2.96	2.92	2.41
	Tibia	2.09	2.12	3.68
MIDDLE BASELINE	**None**	3.00		
TD-BASELINE	**None**	2.60		

means it may be worth revisiting this question as it becomes easier to embed multiple IMU sensors in running apparel.

Table 2: Comparison of the MAE for models learned on all combinations of the four sensor locations: arm, wrist, left tibia, and right tibia. Results for all three learning settings are shown. The models are trained using only the statistical features with a GBRT.

	AM	OM	IM
SENSORS	**MAE**	**MAE**	**MAE**
Arm (A)	1.99	2.03	1.98
Wrist (W)	1.89	2.04	2.15
Tibia (T)	1.98	2.08	2.02
T-T	1.84	1.90	2.10
W-A	1.89	1.95	2.02
T-A	1.98	2.16	1.89
T-W	1.84	2.01	1.98
T-W-A	1.92	1.98	1.97
T-T-A	1.89	2.00	1.96
T-T-W	**1.74**	1.88	2.06
T-T-W-A	**1.83**	1.90	1.99

4.4 Experiment and Results for Q4

Because RPE is a subjective measure, different runners might rate their RPE differently. Therefore, we hypothesized that normalizing the RPE values for training to account for these inter-individual differences will improve predictive performance.

For each learning setting and sensor location, we trained two GBRT models. The first model was trained using the normalized RPE values, like is done in all other experiments in this paper. The second model was trained using the originally reported RPE values. Table 3 reports the MAE for both approaches for each sensor location. Normalizing the RPE clearly improves the MAEs in the AM and OM learning settings. However, when considering Individual Models, there is no real difference between using NRPE and RPE. These results suggest that runners, at least to some extent, consistently report RPE during consecutive tests. However, different people seem to interpret the BORG scale differently. This might be because the runners had no previous experience with the BORG scale. While previous research found that a learning protocol can improve the validity of the BORG scale [27], it is interesting to see that we can account for these subjective differences between runners, as most runners seem to use their warmup RPE as an anchoring point to rate the remainder of the test.

Table 3: The effect of training the model using the normalized RPE (NRPE) values, as is done in all other experiments, and the original RPE values. The normalization is performed to control for the consistency and subjectiveness of the reported RPE values. Results for all three learning settings and each of the three sensor locations are shown. The models are trained using only the statistical features with a GBRT.

	AM		OM		IM	
	NRPE	**RPE**	**NRPE**	**RPE**	**NRPE**	**RPE**
Sensors	**MAE**	**MAE**	**MAE**	**MAE**	**MAE**	**MAE**
Arm (A)	1.99	2.31	2.03	2.38	1.98	2.11
Wrist (W)	**1.89**	2.24	2.04	2.40	2.15	2.12
Tibia (T)	1.98	2.12	2.08	2.29	2.02	1.97

4.5 Experiment and Results for Q5

As the sport science literature has used complex features to study running gait, we hypothesized that these features could complement the set of statistical features and result in improved performance when predicting RPE. Furthermore, we assumed that explicitly describing the symmetry between the movement of the left and the right tibia would capture additional useful information.

In each learning setting, we learned one GBRT model for each combination of feature types: (1) statistical, (2) sports science, (3) statistical and symmetry, (4) sports science and symmetry, (5) statistical and sports science, and (6) statistical, sports science and symmetry. We considered two sensor combinations: wrist (W) and right tibia-left tibia-wrist-arm (T-T-W-A). Note that the symmetry features are only applicable for the second sensor combination.

Table 4 reports the MAE for all the different models. The statistical features alone result in the best or close to the best performance in all three learning settings. There are only small changes in the MAE when considering the more advanced features. These results impact the real-world applicability, as the simple statistical features are computationally less expensive to compute. That is, they can easily be computed in real-time and within the resource constraints of a mobile computing platform worn by a runner.

Table 4: The effect of different combinations of statistical, sports science and symmetry features on the MAE. Results for all three learning settings using the data from the wrist (W) and the combined data from the arm, wrist, left tibia and right tibia (T-T-W-A) are shown. The models are trained using GBRT.

	AM		OM		IM	
Type	**W**	**T-T-W-A**	**W**	**T-T-W-A**	**W**	**T-T-W-A**
Stat.	1.89	**1.83**	2.04	1.90	1.98	1.99
Sport & Symm.	/	1.99	/	2.06	/	2.02
Stat. & Symm.	/	1.84	/	1.97	/	2.16
Sport	2.14	1.92	2.21	2.07	2.09	2.03
Stat. & Sport	1.99	**1.80**	2.05	2.04	2.09	2.01
Stat. & Sport & Symm.	/	1.84	/	1.91	/	1.97

4.6 Experiment and Results for Q6

Both running speed and inter and intra individual differences between runners add noise to the computed feature values. Therefore, we hypothesized that we can improve the prediction of RPE by both (1) standardizing the signals per window before calculating the features and (2) normalizing the feature values with respect to a trial-specific individual baseline for the runner. For each combination of including or excluding these two preprocessing steps, we trained one GBRT model per learning setting using the statistical features calculated on the combined arm, wrist, left tibia and right tibia data.

Table 5 reports the MAE for each combination of the two preprocessing steps. The results indicate that normalizing the feature values with respect to the baseline of a runner is an important step that positively impacts predictive performance. This is in accordance with our hypothesis that running style is highly runner specific. However, the standardization of the signal per window has a limited impact on the results.

Table 5: Impact of standarization and normalization with respect to a trial-specific individual baseline on the MAE. Results for all three learning settings using the combined data from the arm, wrist, left tibia and right tibia are shown. The models are trained using the statistical features and GBRT.

		AM	OM	IM
Standardize Signals	**Normalize w.r.t. Individual Baseline**	**MAE**	**MAE**	**MAE**
yes	yes	**1.83**	1.90	1.99
	no	2.12	2.48	1.95
no	yes	**1.81**	2.00	1.96
	no	2.08	2.50	1.93

4.7 Discussion

We now revisit the questions posed at the beginning of this section. We can positively answer **Q1** as our evaluation showed that our predictive models have a non-trivial performance when predicting RPE while running. Accurate predictions can be made based on a single sensor that could be located on the wrist, arm or tibia, with the wrist yielding the best results (**Q2**). Furthermore, when evaluating **Q3**, we found that fusing data collected from sensors at multiple locations only resulted in slightly improved predictive performance. Somewhat surprisingly, considering advance features coming from the sports science literature (**Q5**) did not result in improved performance compared to only considering standard statistical features. We identified several meaningful preprocessing steps that were important to perform in order to account for both inter and intra individual differences and the subjectivity of the RPE scale (**Q4** and **Q6**). To summarize, it is encouraging that promising results are possible using a single sensor attached to the wrist and a set of computationally efficient features.

In terms of moving more towards deploying such a system "in the wild," considering the impact of external factors such as running surface and weather conditions and internal factors such as individual characteristics and pacing strategies would be important. These factor might, for example, influence the interpolation strategy used to assign an RPE value to each window. Furthermore, exploring the relationship between the accumulated load of the impacts endured while running (both in and across multiple training sessions) and RPE, as has been studied for other sports like professional soccer [13], would be worthwhile.

5 CONCLUSION

This paper introduced fatigue prediction in runners as a new non-trivial, interesting, and impactful data science problem. Specifically, its non-trivial challenges arise from analyzing sensor data collected in an uncontrolled outdoor environment and the need to resort to a subjective partial and evolving truth label for fatigue. More specifically, we showed that the fatigue status of a runner can accurately be predicted with limited or no prior labeled data of a runner using a set of simple features computed on the data of one IMU-sensor attached to the wrist. Moreover, our methodology effectively accounts for running speed, the subjectivity of the target variable and inter and intra individual differences between runners. Thus, the results presented in this work are useful and represent a solid start for moving into a real-world application for monitoring the fatigue level of outdoor runners using wearable sensors.

ACKNOWLEDGMENTS

The authors would like to thank all runners that participated in this study. Tim Op De Beéck and Kurt Schütte are supported by the KU Leuven Research Fund (C22/15/015, C32/17/036). Benedicte Vanwanseele is partially supported by the KU Leuven Research Fund (C22/15/015, C32/17/036) and Interreg V A project NANO4Sports. Jesse Davis is partially supported by the KU Leuven Research Fund (C14/17/070, C22/15/015, C32/17/036), FWO-Vlaanderen (SBO-150033) and Interreg V A project NANO4Sports. The authors have no conflicts of interest to declare.

REFERENCES

[1] M T Ballas, J Tytko, and D Cookson. 1997. Common overuse running injuries: diagnosis and management. *American family physician* 55, 7 (1997), 2473–2484.

[2] G Borg. 1998. *Borg's perceived exertion and pain scales*. Human Kinetics. 1–97 pages.

[3] G A Borg. 1982. Psychophysical bases of perceived exertion. *Med Sci Sports Exerc* 14, 5 (1982), 377–381.

[4] L J Boyd, K Ball, and R J Aughey. 2011. The reliability of MinimaxX accelerometers for measuring physical activity in Australian football. *International Journal of Sports Physiology and Performance* 6, 3 (2011), 311–321.

[5] R M Bryce and K B Sprague. 2012. Revisiting detrended fluctuation analysis. *Scientific reports* 2 (2012), 315.

[6] H Crewe, R Tucker, and T D. Noakes. 2008. The rate of increase in rating of perceived exertion predicts the duration of exercise to fatigue at a fixed power output in different environmental conditions. *European Journal of Applied Physiology* 103, 5 (2008), 569–577.

[7] J Dong. 2016. The role of heart rate variability in sports physiology. *Experimental and Therapeutic Medicine* 11, 5 (2016), 1531–1536.

[8] J H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (2002), 367–378.

[9] D Gafurov, K Helkala, and T Søndrol. 2006. Biometric Gait Authentication Using Accelerometer Sensor. *JCP* 1, 7 (2006), 51–59.

[10] T K Ho. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (1998), 832–844.

[11] S L Hooper and L T Mackinnon. 1995. Monitoring Overtraining in Athletes Recommendations. *Sports Medicine* 20, 5 (1995), 321–322.

[12] A Hreljac, R N Marshall, and P A Hume. 2000. Evaluation of lower extremity overuse injury potential in runners. *Medicine & Science in Sports & Exercise* 32, 9 (2000), 1635–1641.

[13] A Jaspers, T Op De Beéck, M S Brink, W G P Frencken, F Staes, J J Davis, and W F Helsen. 2017. Relationships Between the External and Internal Training Load in Professional Soccer: What Can We Learn From Machine Learning? *International journal of sports physiology and performance* (2017), 1–18.

[14] K Jordan, J H Challis, and K M Newell. 2007. Speed influences on the scaling behavior of gait cycle fluctuations during treadmill running. *Human Movement Science* 26, 1 (2007), 87–102.

[15] Y Kobayashi, T Takeuchi, T Hosoi, H Yoshizaki, and J A Loeppky. 2005. Effect of a marathon run on serum lipoproteins, creatine kinase, and lactate dehydrogenase in recreational runners. *Research Quarterly for Exercise and Sport* 76, 4 (2005), 450–455.

[16] Y Koren. 2010. Collaborative filtering with temporal dynamics. *Commun. ACM* 53, 4 (2010), 89–97.

[17] J R Kwapisz, G M Weiss, and S A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[18] M L Bertelsen, A Hulme, J Petersen, R Korsgaard Brund, H Sørensen, CF Finch, E T Parner, and R O Nielsen. 2017. A framework for the etiology of running-related injuries. *Scandinavian Journal of Medicine & Science in Sports* (2017), 1170–1180.

[19] S Madgwick. 2010. An efficient orientation filter for inertial and inertial/magnetic sensor arrays. *Report x-io and University of Bristol (UK)* 25 (2010), 1–32.

[20] S P Messier, C Legault, C R Schoenlank, J J Newman, D F Martin, and P Devita. 2008. Risk factors and mechanisms of knee injury in runners. *Medicine & Science in Sports & Exercise* 40, 11 (2008), 1873–1879.

[21] R Moe-Nilssen and J L Helbostad. 2004. Estimation of gait cycle characteristics by trunk accelerometry. *Journal of Biomechanics* 37, 1 (2004), 121–126.

[22] J B Morin, P Samozino, and G Y Millet. 2011. Changes in running kinematics, kinetics, and spring-mass behavior over a 24-h run. *Medicine and Science in Sports and Exercise* 43, 5 (may 2011), 829–36. http://www.ncbi.nlm.nih.gov/pubmed/20962690

[23] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[24] J S Richman and J R Moorman. 2000. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology* 278, 6 (2000), 2039–2049.

[25] K H Schütte, E A Maas, V Exadaktylos, D Berckmans, R E Venter, and B Vanwanseele. 2015. Wireless tri-axial trunk accelerometry detects deviations in dynamic center of mass motion due to running-induced fatigue. *PloS One* 10, 10 (2015), e0141957.

[26] K H Schütte, S Seerden, R Venter, and B Vanwanseele. 2016. Fatigue-related asymmetry and instability during a 3200-m time-trial performance in healthy runners. In *ISBS-Conference Proceedings Archive*, Vol. 34. 933–936.

[27] A Soriano-Maldonado, L Romero, P Femia, C Roero, J R Ruiz, and A Gutierrez. 2014. A learning protocol improves the validity of the Borg 6–20 RPE scale during indoor cycling. *International Journal of Sports Medicine* 35, 05 (2014), 379–384.

[28] N M Stoudemire, L Wideman, K A. Pass, C L. Mcginnes, G A. Gaesser, and A Weltman. 1996. The validity of regulating blood lactate concentration during running by ratings of perceived exertion. *Medicine and Science in Sports and Exercise* 28, 4 (1996), 490–495.

[29] J E Taunton, M B Ryan, DB Clement, D C McKenzie, D R Lloyd-Smith, and B D Zumbo. 2002. A retrospective case-control analysis of 2002 running injuries. *British journal of sports medicine* 36, 2 (2002), 95–101.

[30] Y Tochigi, N A Segal, T Vaseenon, and T D Brown. 2012. Entropy analysis of tri-axial leg acceleration signal waveforms for measurement of decrease of physiological variability in human gait. *Journal of Orthopaedic Research* 30, 6 (2012), 897–904.

[31] B RN van Gent, D D Siem, M van Middelkoop, T AG van Os, Sita SMA Bierma-Zeinstra, and B BW Koes. 2007. Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *British journal of sports medicine* (2007), 469–480.

[32] C Wetherell. 1986. The Log Percent (L%): An Absolute Measure of Relative Change. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 19, 1 (1986), 25–26.

[33] T J Williams, G S Krahenbuhl, and D W Morgan. 1991. Mood state and running economy in moderately trained male runners. *Medicine & Science in Sports & Exercise* 23, 6 (1991), 727–31.