Using Biomechanical Signals to Detect Biomechanical Fatigue in Runners using Automated Machine Learning

Michael Mc Cahill and Tomas Ward

michael.mccahill2@mail.dcu.ie

18446652

April 29, 2022

Abstract

Injuries in running have been identified as being one of the most common issues hindering athletic performance. Due to the uptake in recreational running, there has also been an increase in the research carried out around finding the main causes of injuries with the aim of preventing them [10]. The primary goal of this research was to find out the capability of a single inertial measurement unit (IMU) (fitted with an accelerometer, gyroscope, and magnetometer) to differentiate between a runner's biomechanical signature in a fatigued and non-fatigued state across two different running surfaces. Data was captured using IMUs to look at participant's biomechanical signature data. TPOT automated machine learning was employed to classify fatigue in recreational/novice runners. TPOT models were trained on one surface (e.g treadmill) and tested on another surface (e.g tarmac running track) in an attempt to transfer the learning from one running surface to the other to detect participant's biomechanical fatigue. Results show that models often overfit depending on what surface they are trained on. However, when using a mixture of data from both surfaces, fatigue can be identified in athletes (82% accuracy). The results indicate that the biomechanical signature of fatigue is not the same across different surfaces, indicating that each surface requires its own individual fatigue detection model. ***Keywords:*** *Biomechanics, Running Fatigue, Automated Machine Learning (AutoML), Machine Learning*

1. Introduction

Since the start of the Covid-19 pandemic, there has been an increase in the number of people taking up recreational running [4]. The World Athletics Association also stated that during the Covid-19 pandemic that there was a running "boom" [16]. Due to the increase in recreational running, the problem of predicting fatigue becomes an interesting one. Sports scientists are now taking fatigue into account as a risk factor in causing injury [8]. According to studies between

30 % and 75 % of runners are injured annually [20]. If we can predict when the fatigue might occur, then it can aid us in preventing injuries in our runners. By being able to predict fatigue in runners, their workload will be able to be adjusted accordingly. To avoid injury, one must first be able to recognise fatigue and assess its severity. Inertial Measurement Units (IMU) have been used to identify running fatigue successfully [17]. IMUs have become increasingly popular in the classification of human movement. IMUs allow us to identify the body's signals and identify fatigue. The purpose of this study was to determine if it is possible to detect biomechanical fatigue across two different running surfaces using biomechanical signals using data captured by a single IMU sensor fitted with an accelerometer, gyroscope, and magnetometer.

2. Background

The concept of fatigue is a very complex one. Although fatigue is a compound problem in this study it will be considered as a result of taking part in a prolonged strenuous activity [10]. Fatigue is known to cause injuries. Endurance fatigue exposes the body to greater risk if it causes an increase in peak impact accelerations [14]. In [14] treadmill running was used to induce endurance fatigue. Athletes performed box jumps at a fatigued and non-fatigued state and it was found that there was an increased risk of injury when performing box jumps in a fatigued state.

Buckley *et al.* [3] made an effort to distinguish between running in a fatigued state versus running in a non-fatigued state. 21 recreational runners were used as part of this study, where participant's data was analysed using IMUs placed on the right shank, left shank, and lumbar spine. The multistage fitness test (beep test) was used to induce fatigue in participants [11]. Participants reported their perceived level of exertion using the Borg scale [6](*Figure 1*) at each progressive level of the beep test. A Random Forest classifier was used with leave-one-subject-out-cross-validation (LOSOCV) [7] which achieved a 75% accuracy score on a single IMU sensor on the Lumbar Spine. The results from [3] show that a single IMU sensor can be used to distinguish between a fatigued and a non-fatigued run.

Borg Scale Ratings

Exertion Description	**Perceived Exertion Rating**
None text	6
Very, very light	7 to 8
Very light	9 to 10
Fairly light	11 to 12
Somewhat hard	13 to 14
Hard	15 to 16
Hard	15 to 16
Very hard	17 to 18
Very, very hard	19 to 20

Figure 1.

Similarly as part of the research carried out by Holmes, [10] carried experiments to show that running fatigue can be predicted using AutoML. The work done as part of [10] builds on the research conducted in [3] The goal of this paper was to differentiate between non-fatigued running and fatigued running using wearable sensors. The data used in [10] was treadmill running data used from 117 participants who completed a fatiguing run. The biomechanical data comprised of over 2.5-million-foot strikes. Individual IMU signals were explored to predict the biomechanical signature of fatigue in the participants. Fatigue was induced by participants initially running at a self-selected pace for one minute that best represents their typical running pace [5]. The speed of the treadmill was then increased by 2 km/hr every three minutes until a point of exertion was reached. The Rate of perceived exertion was measured using the BORG scale (*Figure 1*). Participants' Rate of Perceived Exertion (RPE) [12] was monitored until a score of 15 was reached on the BORG scale. The TPOT classifier was implemented to classify the participant's run as either fatigued or non-fatigued. Experiments were carried out with and without using a time window. The model's performance was estimated using a 5-fold Stratified cross-validation. F1-score, balanced accuracy and Receiver Operating Characteristics Area Under the Curve (ROC AUC) were used to evaluate model performance. It was found that Gyroscope was the best performing sensor followed by Accelerometer and Magnetometer. The model scores are indicated as being without time window in training and test data as the model was often scoring 100% (portraying model bias), when training data using a time window. The best performing results were a 0.72 balanced accuracy score using a gyrometer for predicting biomechanical fatigue. Holmes [10] states that the model performs better with more data (some cases taking 100% of the available data). This could lead to overfitting of the model when training on another data set. Limitations of the model were noted in [10] as it underperformed using unseen data (track data). This could be due to model bias as stratified cross-validation was used to estimate the model. Overall the results in [10] are promising but the model could be biased as it is learning from previous subject data in the training phase of the model.

Hollis *et al*[9] made an effort to detect biomechanical changes during running between surfaces. The conclusions from this study found that the sensors used in this study were able to detect changes in biomechanical signatures across different surfaces and at different speeds. The sensor used was a Runscribe sensor which is a wearable IMU sensor [26] which provides a detailed view of gait mechanics [1].

AutoML is the process of applying machine learning models to real-world problems using automation [13]. AutoML has become increasingly popular over the last number of years and is continuing to improve results in machine learning projects. TPOT is one of the most widely used AutoML frameworks. TPOT is a tree-based pipeline optimization tool for AutoML [18]. The goal of the TPOT algorithm is to maximise classification accuracy on a supervised classification task. The results from a TPOT framework significantly outperform traditional methods of machine learning classifiers.

The research provided in this paper expands on previous work carried out in [10] and [3] to further investigate the performance of machine learning models to predict fatigue in running. In this study however, a cross-discipline machine learning model will be produced to predict fatigue across two running surfaces (treadmill and tarmac running track). It is mentioned that an attempt to transfer the learning to detect fatigue using data from a different experiment (running on an outdoor track with a beep test used as part of the fatiguing protocol) was encouraging but not as positive. It is also mentioned in [10] that future experiments should consider using a more intense fatiguing protocol. Employing a similar fatigue protocol in [25] where the emphasis was on achieving a condition of extreme fatigue might be more informative when trying to determine the level of fatigue on the biomechanical signature of runners. By investigating biomechanical fatigue across different surfaces, it will be able to be determined if there is a difference in biomechanical signatures across different running surfaces.

3. Methodology

A. Data Sets

Two data sets were used as part of this study to represent athletes running on two different surfaces. Treadmill running data and running track running data were used. Both data sets contained data from a single IMU that generated three raw accelerometer signals, raw gyroscope signals, and raw magnetometer signals. Both data sets are independent to each other and were gathered independently to identify the impact of fatigue on participant in two different ways.

A.1 *Treadmill Data*

This data set comprised of 117 participants (42 females, 75 males) who completed a fatiguing run. All participants identified themselves as novice/recreational runners and hadn't had an injury 3 months before taking part in the experiment [10]. A novice runner in the context of this data set is a runner who sporadically ran up to 10km a week over the past 12 months prior to undergoing participation in the experiment. A recreational runner was defined as a runner that was running at least 10km every week over the last 6 months. Participants were attached with IMUs in the following locations: feet (dorsal aspect of the feet bilaterally, 2 cm above talar dome), shanks (lateral aspect of the shank bilaterally, 5 cm proximal to the lateral malleolus), thighs (lateral aspect of the thighs bilaterally, 10 cm proximal to the lateral knee joint) and one IMU placed directly over the sacrum [10]. To obtain the biomechanical signature of the participants running form, participants ran for one minute at a self-selected pace that best represents their typical running pace. This was used to determine the participant's biomechanical signature in a non-fatigued state. Speed was incrementally increased by 2 km/hr every three minutes until a point of exertion was reached [10]. By incrementally increasing the speed of the treadmill from the runner's initial pace, fatigue was able to be induced in the participants. Participants were monitored every 3 minutes until an RPE of 15 was reached on the BORG scale. Upon completion of the fatiguing run, participants were asked to continue running at their self-selected speed for one minute. Kinematic and kinetic data points were then extracted from the IMUs the initial and exhaustive time point [10]. This data was gathered between January 2018 and September 2019. Data was collected at a frequency of 512 Hz.

Level	Shuttles	Speed (km/h)	Shuttle Time (seconds)	Cumulative Distance (m)	Cumulative Time (min and seconds)
1	7	8	9	140	01:03
2	8	9	8	300	02:07
3	8	9.5	7.58	460	03:08
4	9	10	7.2	640	04:12
5	9	10.5	6.86	820	05:14
6	10	11	6.55	1,020	06:20
7	10	11.5	6.26	1,220	07:22
8	11	12	6	1,440	08:28
9	11	12.5	5.76	1,660	09:31
10	11	13	5.54	1,880	10:32
11	12	13.5	5.33	2,120	11:36
12	12	14	5.14	2,360	12:38
13	13	14.5	4.97	2,620	13:43
14	13	15	4.8	2,880	14:45
15	13	15.5	4.65	3,140	15:46
16	14	16	4.5	3,420	16:49
17	14	16.5	4.36	3,700	17:50
18	15	17	4.24	4,000	18:54
19	15	17.5	4.11	4,300	19:56
20	16	18	4	4,620	21:00
21	16	18.5	3.89	4,940	22:03

Figure 2. Bleep test level description

A.2 *Track Data*

The data used as part of the track data contained data from 21 participants (11 females, 10 males) [3]. All 21 participants were deemed to have been "recreationally active" upon conducting the experiment. Participants had no lower limb injury that would impair their running performance [3]. Participants were attached with IMUs in the following locations: both shanks (2cms above the lateral malleolus) and on the 5th lumbar spinous process. To obtain the biomechanical signature of the participants running form, subjects performed a warm-up run of 400m at a natural running pace that matches their 5km running pace. This was used to determine the participant's biomechanical signature in a non-fatigued state. The beep test was used to induce fatigue in participants. The beep test requires the athlete to perform continuous 20m shuttle runs, whereby the individual must reach the opposite end of the 20m grid before the next beep sounds [15]. After each level of the beep test participants recorded their RPE on the BORG scale. The levels of the beep test can be seen in *Figure 2*.

When a subject could no longer match the pace to reach the next level or when a Borg rating of 18 or higher was reported the test was terminated. A subsequent 400m run was completed within 30 seconds of completing the Beep test (fatigued state) [3]. Data was collected at a frequency of 256 Hz.

B. Data processing

. The data file for each participant was made up of 1.5-2 million data points sampled at a frequency of 512Hz for the accelerometer, gyroscope, and magnetometer for the treadmill data. Similar data points were sampled for the track data at a lower frequency of 256Hz. The accelerometer, gyroscope, and magnetometer contain 3 axes, resulting in 9 axes for every single data point. This meant for each participant there would be over 15 million data points. Instead of analysing all the data, which, given its scale, would have been extremely resource-intensive, a decision was made to focus on two specific time points for both data sets. For the treadmill data, the self-selected pace was used to represent the participant running in a non-fatigued state, and the last 2 minutes of the running protocol to represent running in a fatigued state [10]. For the track data, the self-selected pace was used to represent the participant running in a non-fatigued state, and the 400m fatigued run was used to represent running in a fatigued state. Lastly, participants with missing data points were removed from the study resulting in 0 participants being removed from the track data set and 15 participants being removed from the 117 participants from the treadmill data set.

C. Signal Processing

To reduce the size of both data sets, resampling was performed on the data. Both, data sets were resampled to a frequency of 128 Hz. This was 1/4 the original size of the treadmill data and 1/2 the original size of the track data. Resampling was performed using Scipy's signal library which uses the Fourier method across a given axis [23]. Fourier method breaks down a waveform into alternate representation characterised by the sine and cosine functions of varying frequencies [2].

D. Feature Engineering

Prior to resampling, the data, noise was removed from the data using a Butterworth filter of order n=5. To process the data, feature engineering had to be performed. Time series features were extracted from our data using the Time Series Feature Extraction Library (TSFEL) [27]. By utilising TSFEL it allowed for exploratory feature extraction tasks in three different domains: statistical, temporal, as well as spectral. TSFEL was processed at a time window of 5 seconds for both data sets which was small enough to process the data efficiently, meaning the features were calculated for each 5-second time window [10]. For sampling both data sets at 128 Hz, 1170 features were extracted using time series feature extraction. Two additional features were added which were *"time_window"* to investigate the effect of time with regards to fatigue, as well as a target column that contained the fatigue classification. In the treadmill data, each participant had approximately 70 different time windows recorded. Holmes [10] states it is important to note that some participants had more or less than the expected number of time windows, which is most likely due to the manual data capture. Approximately 35 different time windows were recorded for the track data participants. As this data was also gathered by hand, some participants had more or less than the expected number of time windows. For modelling, data was split using 80% of the data (training data), and 20% of the data was used for testing. Whilst performing tests across both running surfaces, data was standardised using a min-max scaler to bring all feature values to a common scale without distorting the values.

E. Feature Selection

Holmes [10] states that the model used in their study performs better using more data (in some cases taking 100% of the data). By using all the features extracted by TSFEL, the model could be susceptible to overfitting. By using Sklearn's feature importance, the features in the data can be ranked in order of importance i.e. how useful they were at classifying fatigue/non-fatigued states. Mutual information measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency and can be used for feature selection [24]. The 10 features with the highest importance value were used for training and testing the model.

F. Modelling

To find the best classifications for predicting fatigue, TPOT was used. TPOT classifier performs an automated search over a number of machine learning pipelines to find the best possible results for classification [19]. Binary classification was used to establish how effectively each individual IMU could distinguish between a non-fatigued and a fatigued running state. To make sure the TPOT model wasn't picking up subject bias, Leave One Out cross validation was employed. Leave one out cross validation is a special case of cross validation where the number of folds equals the number of instances in the data set. Thus, the learning algorithm is applied once for each instance, using all other instances as a training set, and using the selected instance as a single-item test set [22]. By leaving a group of participants data out each iteration of training the model prevents the same subject data from being split across training and testing data, preventing data leakage. This ensures that every iteration of the model one participant's data is left out from training and testing data. By leaving a participant's data out each iteration of training, the model doesn't pick up any potential subject bias. The behaviour of Leave one out cross validation can be observed in *Figure 3*.

All modelling experiments were carried out with and without the *time_window* feature. It is natural that the longer the participant runs, the more fatigued they are, so by removing the *time_window* feature the classifier can be taught the biomechanical signature of the participants. Models were also run using k-fold cross validation to directly compare results to experiments in [10].

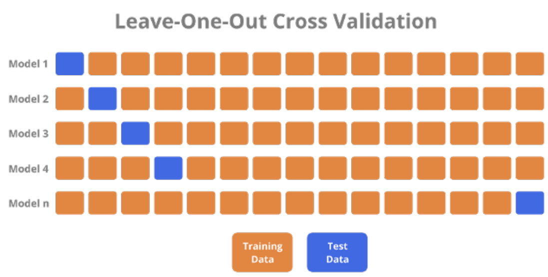

Figure 3. Leave One Out Cross Validation

G. Model Evaluation Metrics

As there isn't one determining metric to evaluate classification, several different metrics were used. Results were compared to results found in [10] using accuracy, F1-Score, balanced accuracy, and Receiver Operating Characteristics Area Under the Curve (ROC AUC). Accuracy was used to show the percentage of times the model classified fatigue correctly. F1-score sums up the predictive performance of the model by combining precision and recall score. In the data, the ratio between non-fatigued and fatigued runs was approximately 5:1. Balanced accuracy was used to address the issue of class imbalance. ROC AUC is used to evaluate how well the model is at distinguishing between classes.

4. Results and Evaluation

Results shown in Tables 1, 2, 3 and 4 are experiments performed without taking time into consideration. Models performed significantly better (Often scoring over 95%) when the length of time they had ran was taken into consideration. To isolate the biomechanical signature of the participants, the time window was removed for the evaluating model performance.

Train: Treadmill Test: Track				
Sensor	**Accuracy**	**F1-Score**	**Balanced Accuracy**	**ROC AUC**
Accelerometer	0.47	0.4	0.35	0.41
Gyroscope	0.48	0.33	0.32	0.39
Magnetometer	0.5	0.35	0.29	035

Table 1.

Train: Track Test: Treadmill				
Sensor	**Accuracy**	**F1-Score**	**Balanced Accuracy**	**ROC AUC**
Accelerometer	0.49	0.48	0.34	0.40
Gyroscope	0.51	0.35	0.43	0.64
Magnetometer	0.49	0.46	0.51	0.55

Table 2.

Track and Treadmill data merged together				
Sensor	**Accuracy**	**F1-Score**	**Balanced Accuracy**	**ROC AUC**
Accelerometer	0.83	0.69	0.66	0.85
Gyroscope	0.84	0.72	0.71	0.79
Magnetometer	0.52	0.34	0.51	0.52

Table 3.

Dual Surface Fatigue Classifier				
Sensor	**Accuracy**	**F1-Score**	**Balanced Accuracy**	**ROC AUC**
Accelerometer	0.74	0.64	0.63	0.74
Gyroscope	0.82	0.64	0.62	0.82
Magnetometer	0.86	0.68	0.64	0.94

Table 4.

Table 1 presents accuracy, F1-score, balanced accuracy, and ROC-AUC for training the TPOT classifier using the treadmill data and testing on the track data. In the training phase of the model, results were high, but the model performance didn't translate across the surfaces.

Table 2 presents accuracy, F1-score, balanced accuracy, and ROC-AUC for training the TPOT classifier using the track data and testing on the track data. Similar to Table 1, results are poor for training the model on one surface and testing on another.

Table 3 presents accuracy, F1-score, balanced accuracy, and ROC-AUC for incorporating all the treadmill and track data together. A random mixture of both data sets is used in this model. Results are promising but due to the size differences in the data sets (treadmill data = 7686 rows, track data = 725 rows), the model is dominated by treadmill data.

Lastly, an experiment was carried out to see how a TPOT model would perform using equal numbers of entries from each data set. Table 4 presents accuracy, F1-score, balanced accuracy, and ROC-AUC for testing the model using a 50:50 surface data split e.g., the model contains 50% treadmill data and 50% track data for training and testing.

Key findings based on results:

- Models overfit for surfaces during the training phase and perform poorly on different surface data.
- Training data isn't representative of the testing data (shows models have overfit).
- Model can predict fatigue using a mixture of data from different surfaces.

The goal of this research was to determine the capability of a single IMU to differentiate between a runner's biomechanical signature in a fatigued and non-fatigued state across two different running surfaces. Having completed the analysis it can be seen that a single IMU is not capable of differentiating between a runner's biomechanical signature in a fatigued and non-fatigued state across two different running surfaces.

5. Discussion

The results that are shown in table 1 and table 2 show that biomechanical fatigue is not transferable across two different running surfaces. There are several contextual factors to this investigation. Fatigue in both instances was induced in a non-natural way (Participants were running with the purpose of becoming fatigued), the running surfaces varied greatly and fatigued was induced in different ways for both data sets. The results show that the concept of biomechanical fatigue is a very complex one and is computed differently across different running surfaces. An example of this would be for Table 1 TPOT determined that the best fit for the model was a random forest classifier and for Table 2 TPOT determined that the best fit for the model was a logistic regression model, for an accelerometer signal. It's important to note that the outcomes of this study could have been influenced by several factors.

A. Limitations

The size and scale of the data sets used in this study were very different. The treadmill data was significantly larger than the track data. The treadmill data was approximately, 900% larger than the track data obtained. Over 80 more participants were involved in the fatiguing studies using the treadmill compared to the track studies. With such a size difference in the data, a true biomechanical fatigue detection system across two different surfaces to determine biomechanical fatigue isn't possible.

It's important to consider that the outcomes of this study could have been influenced by a variety of factors. Running on a treadmill compared to running on a tarmac surface is very different and has different impacts on the body. Treadmills offer better shock absorption than tarmac, which means less stress on the ankles and knees [21]. The way that fatigue was induced in both sets of participants was both different. For the treadmill experiments, fatigue was being determined as the participants were running, meaning fatigue was induced gradually. For the track experiments, fatigue was quickly induced using the beep test and then fatigue was measured after the participant had reached their maximum level on the beep test, rather than it being measured as the participants progressed through the levels of the beep test. The treadmill experiments were also carried out indoors which could also lead to an impact on the results. By having the experiments indoors, the way the experiments are being conducted can be more controlled. The track experiments were carried out on an outdoor running track. It was not mentioned what the weather was like during data acquisition and could have been different for every participant partaking in the experiment. It should also be noted that the experiments relied on the correctness and accuracy of the records made on the day of the test, which were taken manually using a stopwatch, pen, and paper.

There are also limitations when using RPE. One could argue that the BORG scale is not an effective way at recording fatigue. Although it is used extensively in running research and in sports science, it relies on participants' honesty and ability to correctly evaluate their individual level of fatigue. Another issue that stems from using RPE is that it is self-reporting. It is possible that participants indicated fatigue, leading to the fatiguing protocol ending sooner than it would have if a more scientific indicator of fatigue was implemented such as VO2 max.

A lack of computational resources hindered the potential of this study. TPOT couldn't process all features in the data sets without crashing, meaning data wasn't explored at higher frequencies. Data resampling took several hours and couldn't be processed without an upgraded version of Google Colab. Although leave one out is a reliable and unbiased way to estimate model performance, it is a computationally expensive procedure to perform. To run a single generation of TPOT using 10 features and leave one out cross validation, takes a minimum of two and a half hours.

B. Future Work

Future work should consider similar fatiguing protocols across different running surfaces. As both data sets were gathered independently, fatigue was induced differently for both sets of participants. Future experiments should consider employing the same fatiguing protocol on different surfaces to identify if there is a difference in biomechanical fatigue whilst performing the same fatiguing protocols. By employing the same fatiguing protocol, it might be more impactful in creating a more universal biomechanical fatigue detection model.

Both experiments are inducing fatigue to study participants in an artificial way. Both data sets were obtained with the intention of making the participant fatigued. By making a more natural/gradual fatiguing protocol, it would be better suited for use in the real world.

Lastly, when examining features after TSFEL, it is clear that there is an interoperability and explainability issue with the dataset features. Feature names after extraction were abstract and didn't provide a helpful understanding of what they represented. Implementing an explainable artificial intelligence approach would enable the model to be better understood, explainable, interoperable, and transparent for humans. Creating a biomechanical fatigue detection model that was easily understandable by humans would be ground-breaking in this space.

6. Conclusions

The intention of this study was to determine if biomechanical fatigue can be detected using biomechanical signals in runners across two different running surfaces. By conducting a number of experiments using data from treadmill experiments and running track experiments, it can be concluded that biomechanical fatigue cannot be detected across two different running surfaces. Performance metrics shown in the results and evaluation section support the claim that biomechanical fatigue is different on different running surfaces. AutoML model performances perform well on individual surfaces, but they fail to translate performance across surfaces. It is evident that this is a problem that still needs investigating as using the same amounts of data from different surfaces yields promising results.

References

[1] *Abnormal gait — walking problems*. Mar. 2022. URL: `https://medlineplus.gov/walkingproblems.html`.

[2] Pete Bevelacqua. *Fourier transforms*. URL: `https://www.thefouriertransform.com/`.

[3] Cillian Buckley et al. "Binary classification of running fatigue using a single inertial measurement unit". In: *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE. 2017, pp. 197–201.

[4] Marley Dickinson. *Strava year-End Report: Runners logged 800 million more kilometres in 2021*. Dec. 2021. URL: `https://runningmagazine.ca/the-scene/strava-year-end-report-runners-logged-800-million-more-kilometres-in-2021/`.

[5] Tracy A Dierks, Irene S Davis, and Joseph Hamill. "The effects of running in an exerted state on lower extremity kinematics and joint timing". In: *Journal of biomechanics* 43.15 (2010), pp. 2993–2998.

[6] Mohamed Elshafei and Emad Shihab. "Towards detecting biceps muscle fatigue in gym activity using wearables". In: *Sensors* 21.3 (2021), p. 759.

[7] Tadayoshi Fushiki. "Estimation of prediction error by using K-fold cross-validation". In: *Statistics and Computing* 21.2 (2011), pp. 137–146.

[8] Youlian Hang. "Routledge Handbook of Ergonomics in Sport and exercise". In: *Routledge Handbook of Ergonomics in Sport and Exercise* (2013). DOI: `10.4324/9780203123355`.

[9] Christina R Hollis et al. "Running biomechanics as measured by wearable sensors: effects of speed and surface". In: *Sports biomechanics* (2019).

[10] Monica Holmes. *Fatigue Detection in Treadmill Running using Wearable Sensors*. Dublin City University, 2021.

[11] Luc A Leger and J_ Lambert. "A maximal multistage 20-m shuttle run test to predict VO2 max". In: *European journal of applied physiology and occupational physiology* 49.1 (1982), pp. 1–12.

[12] Sara Lindberg. *RPE: What does this scale tell you about exercise?* Mar. 2019. URL: https://www.healthline.com/health/RPE.

[13] Ben Lutkevich. *What is Automated Machine Learning (automl)?* Oct. 2020. URL: https://www.techtarget.com/searchenterpriseai/definition/automated-machine-learning-AutoML.

[14] Kieran A Moran et al. "Does endurance fatigue increase the risk of injury when performing drop jumps?" In: *The Journal of Strength & Conditioning Research* 23.5 (2009), pp. 1448–1455.

[15] *Multistage Fitness (BEEP) test*. July 2020. URL: https://www.scienceforsport.com/multistage-fitness-beep-test/.

[16] *New research reveals running boom during covid-19 pandemic: Press-releases: World Athletics*. URL: https://www.worldathletics.org/news/press-releases/global-running-day-research-nielsen.

[17] Martin O'Reilly et al. "Binary Classification of Running Fatigue using a Single Inertial Measurement Unit". In: May 2017. DOI: 10.1109/BSN.2017.7936040.

[18] Randal S. Olson and Jason H. Moore. "TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning". In: *Proceedings of the Workshop on Automatic Machine Learning*. Ed. by Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Vol. 64. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, June 2016, pp. 66–74. URL: https://proceedings.mlr.press/v64/olson_tpot_2016.html.

[19] Randal S. Olson et al. "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science". In: *Proceedings of the Genetic and Evolutionary Computation Conference 2016*. GECCO '16. Denver, Colorado, USA: ACM, 2016, pp. 485–492. ISBN: 978-1-4503-4206-3. DOI: 10.1145/2908812.2908918. URL: http://doi.acm.org/10.1145/2908812.2908918.

[20] Alvin Powell. *Where runners go wrong*. Mar. 2019. URL: https://news.harvard.edu/gazette/story/2016/02/where-runners-go-wrong/.

[21] Elizabeth Quinn. *Is it better to run on a treadmill or outside?* May 2021. URL: https://www.verywellfit.com/treadmill-vs-outside-running-which-is-best-3120796#:~:text=Treadmills%20offer%20better%20shock%20absorption,be%20hard%20on%20your%20body..

[22] "Leave-One-Out Cross-Validation". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 600–601. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_469. URL: https://doi.org/10.1007/978-0-387-30164-8_469.

[23] *Scipy.signal.resample*. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.resample.html.

[24] *Sklearn.feature$_s$election.mutual$_i$nfo$_c$lassif*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html#sklearn.feature_selection.mutual_info_classif.

[25] Oleg Verbitsky et al. "Shock transmission and fatigue in human running". In: *Journal of applied biomechanics* 14.3 (1998), pp. 300–311.

[26] *Wearable IMU - gait analysis*. June 2019. URL: https://runscribe.com/.

[27] *Welcome to TSFEL documentation!* URL: https://tsfel.readthedocs.io/en/latest/.