Small Data, Big Challenges: Pitfalls and Strategies for Machine Learning in Fatigue Detection

André Jeworutzki
University of the West of Scotland
Hamburg University of Applied Sciences
Hamburg, Germany
andre.jeworutzki@haw-hamburg.de

Jan Schwarzer
Hamburg University of Applied Sciences
Hamburg, Germany
jan.schwarzer@haw-hamburg.de

Kai von Luck
Hamburg University of Applied Sciences
Hamburg, Germany
kai.vonluck@haw-hamburg.de

Peer Stelldinger
Hamburg University of Applied Sciences
Hamburg, Germany
peer.stelldinger@haw-hamburg.de

Susanne Draheim
Hamburg University of Applied Sciences
Hamburg, Germany
susanne.draheim@haw-hamburg.de

Qi Wang
University of the West of Scotland
Paisley, Scotland
qi.wang@uws.ac.uk

ABSTRACT

This research addresses the pitfalls and strategies for machine learning with small data sets in the context of sensor-based fatigue detection. It is shown that many existing studies in this area rely on small data sets and that classification results can vary considerably depending on the evaluation method. Our analysis is based on a study with 46 subjects performing multiple sets of squat exercises in a laboratory setting. Data from ratings of perceived exertion, inertial measurement units, and pose estimation were used to train and compare different classifiers. Our findings suggest that commonly used evaluation methods, such as leave-one-subject-out, should be used with caution and may not lead to generalizable classifiers. Furthermore, challenges related to imbalanced data and oversampling are discussed.

CCS CONCEPTS

• **Human-centered computing → Laboratory experiments**.

KEYWORDS

small data, imbalanced data, oversampling, model evaluation, class distribution, fatigue detection, machine learning, pose estimation, IMU, wearable sensors, sports, exercise, squats, RPE

ACM Reference Format:
André Jeworutzki, Jan Schwarzer, Kai von Luck, Peer Stelldinger, Susanne Draheim, and Qi Wang. 2023. Small Data, Big Challenges: Pitfalls and Strategies for Machine Learning in Fatigue Detection. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594806.3594825

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PETRA '23, July 05–07, 2023, Corfu, Greece

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0069-9/23/07…$15.00
https://doi.org/10.1145/3594806.3594825

1 INTRODUCTION

Fatigue[1] increases the risk of injury and reduces exercise performance [37, 61]. Early detection of fatigue during exercises would help adjust training to prevent over-training and injury [33, 61]. According to Wang et al. [67], machine learning has been widely used by researchers for health care and activity monitoring.

The aim of our study is to identify and address the pitfalls and strategies for machine learning in sensor-based fatigue detection with small data. By analyzing evaluation methods and the impact of small data sets on classification results and exploring strategies for handling imbalanced data and improving generalization, we hope to provide insights that can help improve the effectiveness and reliability of fatigue detection systems. *Small data* in this context is understood as data from a small number of different people to achieve individual-level prediction [31][2]. Toward this aim, we conducted a literature review as well as our own study with 46 subjects to train classifiers that detect three fatigue levels during squat exercises based on *ratings of perceived exertion* (RPE), *inertial measurement units* (IMU) and *pose estimation* (PE). Our research is guided by related studies, e.g., [1, 4, 15, 18, 21, 22, 28, 33, 36, 37, 51, 66, 67]. Most of them achieved classification accuracy of at least 80% with similarly small data sets. Compared to related studies, we show that the classification results significantly depend on both the evaluation method and the way data from different subjects are handled, especially in terms of imbalanced data, class distribution, and oversampling. Although our analysis is based on fatigue detection for squats with data from IMU and PE, our results should also apply to other sensors and exercises.

This document is organized as follows: The related literature is presented in Section 2 including an outline of the review methods. Section 3 elaborates the research methods including the research setting, sample selection, and procedure for data collection and analysis. Section 4 illustrates the results of our study. Section 5 discusses the results, limitations, and potential pitfalls when dealing

[1] A detailed discussion on the definition of fatigue can be found in [2, 54]. In this paper, *fatigue* is understood as the feeling of physical exhaustion caused by physical activity.

[2] Several notions for small data exist, see for example Faraway and Augustin [23], Kitchin and Lauriault [38], Thinyane [64].

with different evaluation types, imbalanced data, and oversampling. Section 6 concludes with recommendations for future work.

2 RELATED WORKS

As a basis for our own study, we conducted a literature review to obtain an overview of common techniques for detecting fatigue during physical activity. We were particularly interested in the evaluation methods, but also in what was used as ground truth and how imbalanced data was handled. For this purpose, we applied the following search phrase in academic search engines (ACM, IEEE, Google Scholar, Scopus):

```
fatigue AND (sports OR training) exercise AND
(machine OR deep) learning
```

For each search engine, the abstracts of the 100 most recent studies were reviewed to determine whether wearables or cameras were used to detect fatigue and whether fatigue was caused by a specific physical activity. Studies that did not use any type of machine learning were excluded. The remaining 38 studies (see Table 1) recruited different numbers of healthy subjects, ranging from 4 to 60, with an average of 19.4 subjects[3]. A wide range of sensors, exercises, and machine learning models are used to detect different levels of fatigue:

Wang et al. [67] investigated whether IMUs attached to pelvis, thigh, and shank are able to predict three fatigue levels with 19 habitual runners. They applied random forest and *support vector machine* (SVM) classifiers and achieved 91.1% accuracy. Chen et al. [20] utilized electromyography sensors at the biceps and brachioradialis in 10 subjects during dumbbell exercises. A *long-term short-term memory* (LSTM) *neural network* (NN) achieved 90.4% accuracy in discriminating between fatigue and non-fatigue. Guo et al. [29] classified three fatigue levels based on blood flow and heart rate during biceps curls with 10 subjects. Their custom-built model achieved 92% accuracy, while a SVM model achieved 83%, and a NN model achieved 88%. Chalitsios et al. [18] determined fatigue using a force plate and IMU from 13 runners on a treadmill. Fatigue was determined by reaching a ventilation threshold. Using a random forest model, they achieved 91.4% accuracy. Zhu et al. [73] detected six fatigue levels depending on six exercises: lying, sitting, walking, 2x cycling, and running. They utilized a *convolutional neural network* (CNN) to classify an electrocardiogram signal of 24 subjects with an accuracy of up to 97.7%. Zhang and Wang [71] determined fatigue and non-fatigue states in 20 athletes based on eyelid closure: a general model first learned the individual characteristics of an athlete to build a more accurate, adaptive fatigue detection model. Depending on the athlete, they achieved between 80% and 90% accuracy with a SVM model. Wang and He [66] determined four levels of fatigue in 12 subjects using depth images from two Kinect sensors during treadmill trials. Depending on the model and the type of training, accuracy values between 62.91% and 90.19% were achieved. Triantafyllopoulos et al. [65] distinguished 14 fatigue levels in 48 runners with audio signals from a smartphone and obtained a mean absolute error of 2.35 with a CNN model. Jiang et al. [33] used a combination of multiple IMUs and a force plate to predict ten fatigue levels during various exercises (squats, high knee lifts, and corkscrew toe touches) from 14 subjects. They applied random forest and CNN models and achieved accuracy values between 89% and 94%. Shi et al. [61] assessed 10 subjects using a combination of heart rate, oxygen consumption, and hip joint angle to determine five fatigue levels during treadmill exercise. They trained seven different machine learning models that achieved accuracy values ranging from 38 to 89%. Overall, the majority of the reviewed studies report comparatively small data sets and achieved accuracy results of at least 80% (see Table 1).

2.1 Ground Truth

The majority of studies (63.16%) are based on subjective RPE as ground truth (see Table 1) – especially Borg [11] or similar adapted scales. The RPE values are repeatedly queried from a subject, mostly based on time, e.g., every 30 seconds [1], every minute [28], or every two minutes [36, 58, 66]. Milanez et al. [47] collected RPE values every ten minutes as well as for the whole training session. Another approach is to collect RPE after certain number of repetitions, e.g., every five repetitions [34]. Alternatively, some studies rely on the intensity of the exercise as ground truth to label fatigue levels, e.g., Zhu et al. [73] associate six levels of fatigue to six exercises of increasing intensity. Chalitsios et al. [18] assessed fatigue when a ventilatory threshold was reached. Cheah et al. [19] used a time-based threshold and labeled the last 20% of repetitions as fatigue.

The number of fatigue levels (i.e., classes) range from 2 to 14. Binary classification is applied by 60.53% of the studies (see Table 1)[4]. The remaining studies use multiple classes to varying degrees.

2.2 Imbalanced Data

Imbalanced data occurs when data sets have a skewed distribution of classes that reduces the classification accuracy [62, 63]. According to Spelmen and Porkodi [62], the main causes for imbalanced data are the lack of density in the training data set, the presence of small disjuncts, the overlapping between classes, the identification of noisy data, the significance of the borderline instances, and the data shift between the training and the test distributions.

Methods exist to deal with imbalanced data, for example, balanced accuracy [14] makes low performance of minority classes visible, certain loss functions like the focal loss [44], or class weights [74] decrease the influence of majority classes. Imbalance can also be reduced directly by undersampling the majority classes or oversampling the minority classes [62]. For small data sets, undersampling is not an option, as it further reduces the size of the data set. The most common problem encountered with oversampling is that no new information is added to the data set, which can lead to overfitting [62]. SMOTE is the most commonly used oversampling technique to create artificial data [12, 24, 28, 55, 62][5].

Only a few studies report how the data of each class is distributed and how imbalanced data is handled. For example, Jiang et al. [34] duplicated collected observations. Guan et al. [28] applied SMOTE to increase the data of the minority classes. Aguirre et al. [1] mapped

[3] Martins et al. [46] presented a broad review on fatigue not limited to machine learning and training exercises. They found similar numbers, with studies involving 3 to 50 subjects, with an average of 14 subjects. Marotta et al. [45] reviewed studies on fatigue based on accelerometers and the lower limbs during cyclical physical exercise and account for 3 to 222 subjects per study with an average number of 23.1 subjects.

[4] E.g., Elshafei et al. [21] consider RPE values >16 as fatigue on the Borg scale.

[5] Other generative models exist such as GANs, normalizing flows, or diffusion networks [10], but they require sufficiently large data.

Table 1: Overview of the related works.

Ref.	Year	n	Data	Exercises	Ground Truth	Classes	Classifiers	Test (accuracy)
[65]	2022	48	Audio, HR, Knee, Foot	Running	RPE20	14	CNN	2.35 (MAE)
[73]	2022	24	HR	Lie, Sit, Walk, Cycle, Run	Exercise intensity	6	CNN	97.7%
[42]	2022	20	HR, EMG	Pilates	RPE	3	SVM, NN, *k*-NN	87.83%
[67]	2022	19	IMU	Running	RPE20	3	SVM, RF	91.1%
[18]	2022	13	VO2, IMU, Force-plate	Treadmill	Ventilatory threshold	2	RF	91.4%
[34]	2022	12	IMU	Squats	RPE10	2	DeepConv LSTM / CNN	83.7%
[61]	2022	10	HR, VO2, Angle	Treadmilll	Exercise intensity	5	LR, SVM, DT, RF, NN	38–89%
[29]	2022	10	Blood Flow, HR	Bicep Curls	RPE	3	SVM, NN, gcForest	92%
[22]	2022	7	IMU	Assembly work	RPE delta	4	NN, SVM, RNN	up to 96.1%
[19]	2022	4	EMG, IMU	Sit-ups	Last 20% segments	2	Gaussian Classifier	65,3% (Avg. F_1)
[53]	2022	N/A	Existing data set	Throwing	N/A	N/A	LSTM	Correlation Coefficient
[1]	2021	60	HR, Kinect	Sit-to-Stand	RPE10	3	RF, LR, DT, *k*-NN, SVM, NB, NN, others	82.5%, 82.7% (Avg. F_1)
[37]	2021	24	IMU	Gait	RPE	2–4	SVM	91%, 78%, 64%
[41]	2021	24	IMU, HR	Material Handling	RPE	2	LSTM Recurrent NN	65% accuracy
[21]	2021	20	IMU	Bicep Curls	RPE20	2	LR, RF, DT, NN	Various
[33]	2021	14	IMU	Squats, Jacks, Toe-touch	RPE10	10	RF, CNN	89–94%
[28]	2021	14	HR, IMU	Treadmill	RPE20	3	Bi-LSTM, SVM, RF, NN	80.55%
[5]	2021	14	IMU	Running	Time intervals	2	LSTM	Silhouette Score
[66]	2021	12	Kinect	Treadmill	RPE	4	CNN, RF, DT, SVM, NN	87.69–90.19%
[20]	2021	10	EMG	Dumbbell	Exhaustion	2	LSTM	90.4%
[40]	2021	9	IMU	Material Handling	RPE10	5	SVM	up to 83.8%, 80.9% (F_1)
[59]	2020	24	IMU, HR	Material Handling	RPE	2	RF, SVM, LR	>85%
[71]	2020	20	Camera / Eye Lid	N/A	Already labeled data set	2	SVM	80–90%
[27]	2020	13	IMU, Camera	Gait	RPE10	2	DT, NB, SVM, *k*-NN	84.62%
[48]	2020	13	EMG	Trunk Exercises	N/A	2	CNN	by Percentage Error
[56]	2020	8	HR	Material Handling	RPE20	2	*k*-NN	up to 78.18%
[50]	2020	8	HR	Material Handling	RPE20	2	SVM, *k*-NN, DT, NB, LR, RF, NN, other	up to 90.36%
[36]	2019	24	IMU	Gait	RPE	2, 4	SVM	91%, 61%
[68]	2019	20	EEG	Muscle Chair	RPE10, Subject Fatigue Scale	2	SVM	up to 90%
[52]	2019	10	EMG	Shoulder + Elbow	RPE	2	SVM, RF, others	77.8% (Avg. F_1)
[35]	2018	52	EMG	Biceps Curls	EMG signals	2	SVM, NB, RF, others	91.5%
[51]	2018	29	IMU, HR	Running	RPE20	14	NN, Regression Trees	MAE
[4]	2018	20	IMU	Gait + Material Handling	RPE, Subjective Fatigue Level	2	SVM (majority voting)	90%
[32]	2018	20	EMG, HR	Treadmill	HRmax	2	NB	98%
[26]	2018	N/A	IMU, HR	Walking, Running, Skiing	Cluster comparison	N/A	NN, Regression	MAE
[15]	2017	21	IMU	Stride	RPE, Beep / Pacer Test	2	RF, SVM, *k*-NN, NB	75%, 75% (F_1)
[9]	2015	31	EMG, MMG	Treadmill	Bruce Protocol	2	NN	92%
[72]	2013	17	IMU	Gait	A trial for each class	2	SVM, various kernels	90%

decision tree (DT), electroencephalogram (EEG), logistic regression (LR), naive bayes (NB), random forest (RF), k-nearest neighbours (k-NN), convolutional neural network (CNN), electroencephalogram (EEG), electromyography (EMG), heart rate (HR), long-short term memory (LSTM), mean absolute error (MAE), mechanomyogram (MMG), neural networks (NN), oxygen consumption (VO2), support vector machine (SVM)

each class to the five closest repetitions for each subject. Jiang et al. [34] divided the subjects into fast- and slow-tiring sub-groups according to the number of repetitions they have conducted prior to exhaustion. Baghdadi et al. [4] extracted data from the first and last ten minutes to obtain an equal number of strides in both fatigued and non-fatigued states.

2.3 Evaluation Types

The reviewed studies use different approaches to evaluate trained classifiers. Wang and He [66] employed three types to test the performance of machine learning models:

- *Type 1*: data of individuals is used to predict individuals (e.g., if multiple data sets of one person exist).
- *Type 2*: data of all subjects are disordered, and a portion is randomly selected as test set, and the rest as training set[6].
- *Type 3*: data of a subject is used as the test set and excluded from the training/validation set (leave one subject out [7]).

Each of these types is usually combined with cross-validation [1, 4, 15, 21, 27, 33, 34, 37, 42, 51–53, 66–68], in which the training set is partitioned into k parts. In each iteration, the model is then trained on k-1 parts and validated against the unknown remaining part. The partitioning can be done in such a way that either all possible permutations or only certain subsets are tested. Cross-validation can also be performed with the test set, for example, in *Type 3*, the test subject is swapped with another subject and the trained classifier is tested again until all subjects (or a limited, random number) have been the test subject once. Subsequently, either the result of the classifier with the best result is picked [28] or an average value is calculated over all performed tests [52, 68]. Depending on the total number of subjects, a single subject for evaluation usually results in a low test rate. To compensate for this problem, we propose a fourth type based on our analysis to reduce the dependence of the trained classifier on one subject and also increase the size of the test set relative to the training set (see also [9, 59]):

- *Type 4*: data of multiple subjects is used as test set and excluded from the training set (leave n subjects out).

Type 4 can be combined with cross-validation as well, in which multiple subjects are selected as test set and multiple test sets are created for cross-validation. To reduce the number of test sets, only a limited number of (random) test sets may be used (Monte Carlo cross-validation [42]).

A further approach is to train multiple classifiers, each with a different test set: the final decision is made by majority vote, which also reduces the dependency on a specific test set or subject(s) [4].

3 METHOD

This section describes our study, which was conducted in 2022. The purpose of this study was to investigate the effects of the identified evaluation types and data augmentation techniques on

[6] A training set is usually further divided into a training and validation set.

the classification results. Our study was primarily guided by the works of Jiang et al. [33, 34], Shi et al. [61], Wang et al. [67].

The structure of this section is as follows: Firstly, the research setting is presented; secondly, the sample selection is described; thirdly the method for data collection is elaborated; and fourthly, the process of data analysis is explained.

3.1 Research Setting

The study took place on multiple weekdays between 10:00 and 15:00 in a laboratory (see Figure 1), where the main author is employed as research assistant. An Ethical approval for the study was granted by the University of the West of Scotland Ethics Committee. The experimental setup consisted of cameras and wearable sensors: a custom-made IMU based on Bosch BMI160 with 200 Hz sampling rate, two c920 webcams by Logitech International S.A. to record each subject from the front and left side during the exercise with 30 Hz and 720p resolution, and infrared cameras by ART GmbH & Co. KG for verification purposes.

3.2 Sample Selection

As described in Section 2, related studies recruited varying numbers of healthy subjects, with an average of 19.4 subjects. A similar cohort was targeted for this research. In the process, the number of subjects was continuously expanded to examine the effect on classification: 20 subjects in the first, 10 in the second, and 16 in the third cohort, for a total of n=46 (31 males and 15 females).

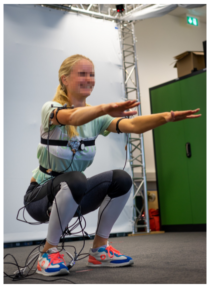

Figure 1: Exercises for squatting in the laboratory.

Morris et al. [49] point out that variations inevitably affect recognition accuracy and thus encourage large-scale training. However, when large-scale training is not possible, it is crucial to reduce the amount of variations. Thus, a balanced set of training data was targeted to avoid class imbalances that would affect classification performance [63]. For this reason, healthy volunteer students from different non-sport disciplines with similar fitness level (occasional weekly fitness routine) and age (between 20 and 30 years) were recruited to obtain a homogeneous group.

3.3 Data Collection

Data was collected in sessions. Each subject performed a session that consisted of three sets of squats[7] with two breaks in between (see Figure 2). Before the start of a session, the subject was informed about the study and asked to sign a consent form. After signing, the subject was equipped with sensors; the IMU was attached near the sternum. An explanation of the procedure and demonstration of the exercise followed. A script was run to start and approximately synchronize all sensors. The subject began the session and performed three consecutive sets of squats. Sensor data was collected throughout the session, including breaks. The subject was asked to state a rating of perceived exertion from 6 to 20 on the Borg RPE Scale [11] every 10 seconds during exercise. Each exercise took 60 seconds followed by a break of 60 seconds. A session, including the introduction, took about 10 minutes per subject.

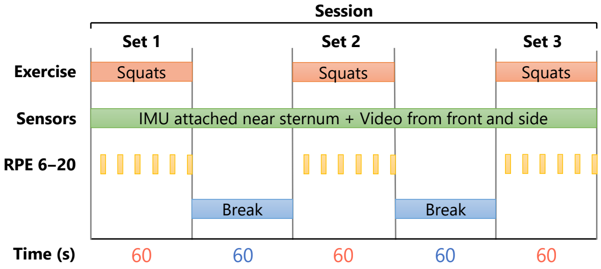

Figure 2: Laboratory protocol for three sets of squats.

3.4 Data Analysis

A commonly used general-purpose framework to design and evaluate activity recognition systems is the Activity Recognition Chain [60] which was introduced by Bulling et al. [17] and consists of five steps: raw data, pre-processing, segmentation, feature extraction, and classification. The following describes how the five steps were implemented to train classifiers that determine fatigue levels with IMU or PE data:

In the first step, raw data is collected from sensors. The sensors are described in Section 3.1.

In the second step, the collected data was pre-processed. In line with related studies (e.g., [30]), the Euclidean norm was applied to the acceleration and gyroscope signals of the IMU to combine the respective x-, y-, and z-axes. In doing so, an exact orientation of the body-worn IMU is no longer required and the number of features can be reduced (see the fourth step). In addition, MediaPipe Pose[8] was utilized with default parameters to extract 32 landmarks (joint coordinates) from both video cameras [6]. The landmarks for fingers and toes were discarded afterwards. Joint angles between two neighboring landmarks were calculated. The angular and joint velocity were then calculated for each landmark based on two consecutive frames (similar to [1]) to ensure that the data between subjects is independent of body proportions and thus comparable.

In the third step, all pre-processed sensor data was segmented into logical units that represent a complete cycle for performing a squat (i.e., repetition). The landmark coordinates of the left pelvis joint were used as the basis for searching for segments, since the peaks on the y-axis (upward and downward movement) can be easily identified [1]. The literature indicates different means to accomplish repetition detection such as minima and maxima searches [43], also known as Zero-Velocity Crossing [13]. Zero-Velocity Crossing is prone to oversegmentation (too many detected segments) [13],

[7] Squats were chosen to induce fatigue because they can be performed multiple times for one minute by most healthy people (compared to, e.g., push-ups).

[8] https://google.github.io/mediapipe/solutions/pose.html

but it has shown to be sufficient for repetitive training exercises in our initial tests. To facilitate peak detection, a third-order low pass Butterworth filter [8] with cut-off frequency of 20 Hz was applied to removed noise and outliers from the signal; the phase shift was corrected by subtracting a constant value. Since the number of repetitions was counted manually, a script could adjust the parameters of the peak detection algorithm until the expected number of peaks was found. The MATLAB findpeaks function was utilized for this purpose (see Figure 3). The first and last segments were then removed because they commonly have additional motions that distinguish them from the rest [39].

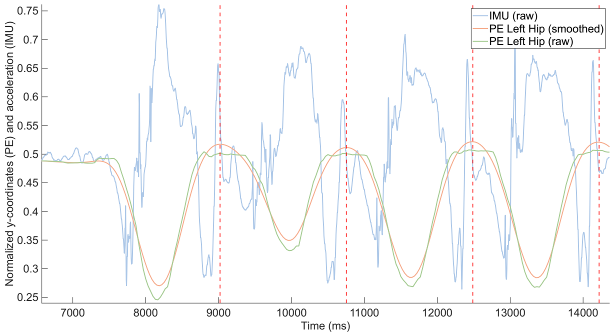

Figure 3: Segmentation based on PE. First segment omitted. The signals are normalized for visualization purposes.

In the fourth step, features were calculated for each segment. Different types of features exist for such a task, like dynamic [8], statistical [30], or frequency-based features [3]. The features used in this study were largely inspired by Guo et al. [30]: skewness, kurtosis, standard deviation, variance, mode, median, range, root mean square, 25th percentile, 75th percentile, and the duration of a segment. These features were applied on the Euclidean values of the acceleration and gyroscope signals as well as the angular and joint velocities. A n-dimensional feature vector was created consisting of a summary of calculated values (e.g., mean, median, and variance) normalized from 0 to 1. The feature vectors have 23 dimensions for the IMU, 67 dimensions for left-side PE, and 122 dimensions for front-side PE.

The RPE values were added to each corresponding feature vector (i.e., segment) based on the timestamps. Since RPE was collected every 10 seconds, the values were linearly interpolated[9] and rounded to obtain RPE values for the segments between the 10-second intervals (see also [1]). RPE values were mapped (6–10, 11–14, and 15–20) to reduce the number of classes from 14 to 3. This assignment was chosen to approximate the original data distribution (see Figure 4) and to improve classification results [40]; experiments revealed that more classes would decrease the overall performance of the classifiers. Guan et al. [28] used the following three thresholds: 6–11, 12–16, and 17–20.

SMOTE oversampling without extensions [24] was then utilized to approximately balance the number of segments between classes: For each segment, another segment of the same class and subject was found by k-NN search (k=3) and used to generate a new segment [24]. This way, 500 (13.91%) new segments were generated for RPE class 2. The total number of segments was 4095. Figure 4 shows the respective distributions.

Table 2: List of classifiers and abbreviated parameters.

Classifier	Parameter 1	Parameter 2	Parameter 3
Binary Tree	SplitCriterion = gdi	MaxNumSplits = 100	Surrogate = off
Bagged Trees	MaxNumSplits = 1002	LearningCycles = 30	
k-NN	Distance = Euclidean	NumNeighbors = 5	DistWeight = Equal
Subspace *k*-NN	Dimension = numFeatures / 2	LearningCycles = 30	
SVM	KernelFunction = gaussian	KernelScale = 9.8	BoxConstraint = 1
NN	DenseLayerSizes = 25	Activations = relu	IterationLimit = 1000

In the fifth step, classifiers were created based on the feature vectors. Each classifier determined fatigue levels for each segment. In the literature, different machine learning techniques are employed (see Table 1). Most of the studies leveraged and compared multiple classifiers, e.g., [1, 15, 21, 33, 35, 42, 59, 61, 66, 67]. However, there is no clear consensus as to which classifier performs best.

Our research was designed to create classifiers for detecting fatigue in the general population. Therefore, only *Type 2–4* (see Section 2.3) were evaluated. Table 2 shows all classifiers and their parameters. Each classifier was trained with 5-fold cross-validation [25]. For *Type 2*, all classifiers were trained with 85% random segments as training set and the remaining 15% was used as test set with cross-validation. For *Type 3*, the test set always consisted of all segments of one subject with subject-based cross-validation. For *Type 4*, as many subjects were included in the training set until 15% was reached for the test set. To ensure that the results were independent of a particular test set, ten random test sets were generated for *Type 4* (Monte Carlo cross-validation). After completion of the cross-validation for the test set, all accuracy and F-score [57] results were averaged.

4 RESULTS

This section describes the classification results. Figure 5 displays how the average accuracy and F_1 score develop over an increasing number of subjects for different classifiers. As the number of subjects increases, average accuracy and F_1 scores tend to decrease across all evaluation types and classifiers. For *Type 2*, the average F_1 scores are often similar to their corresponding average accuracy values. *Type 2* also achieved the highest average accuracy values and F_1 scores overall. For *Type 3 and 4*, average accuracy and F_1 score vary more strongly and are more often apart from each other. For *Type 4*, the mean values for average accuracy and F_1 score are similar to *Type 3* values. The mean values between IMU and PE Side are often similar, while average F_1 score is usually lower than average accuracy.

Table 3 compares how *Type 2* performs with data from unknown subjects. For this purpose, new *Type 2* classifiers were trained with data from 39 of the 46 subjects. The remaining subjects were used as the unknown test set (~15% test data). The average accuracy and F_1 score decrease up to 42.43% percentage points when *Type 2* classifiers are confronted with data from unknown subjects whose data is not present in the training set (a similar trend can be observed when 45 subjects are used for the training set and one subject for the test set).

Figure 6 compares how different classifiers perform with different evaluation types and feature sets ("Combined" means that all features from IMU, PE Front, and PE Side were combined). *Type 2* achieves the highest average accuracy and F_1 score values,

[9] RPE is known to linearly change with exercise intensity [11, 51].

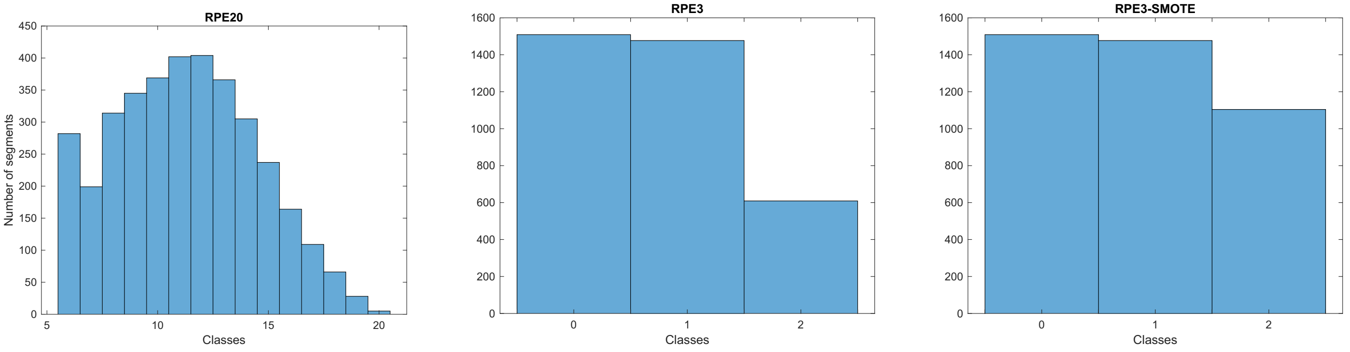

Figure 4: Distribution of segments by RPE classes. Left: 14 classes. Middle: 3 classes. Right: 3 classes including SMOTE.

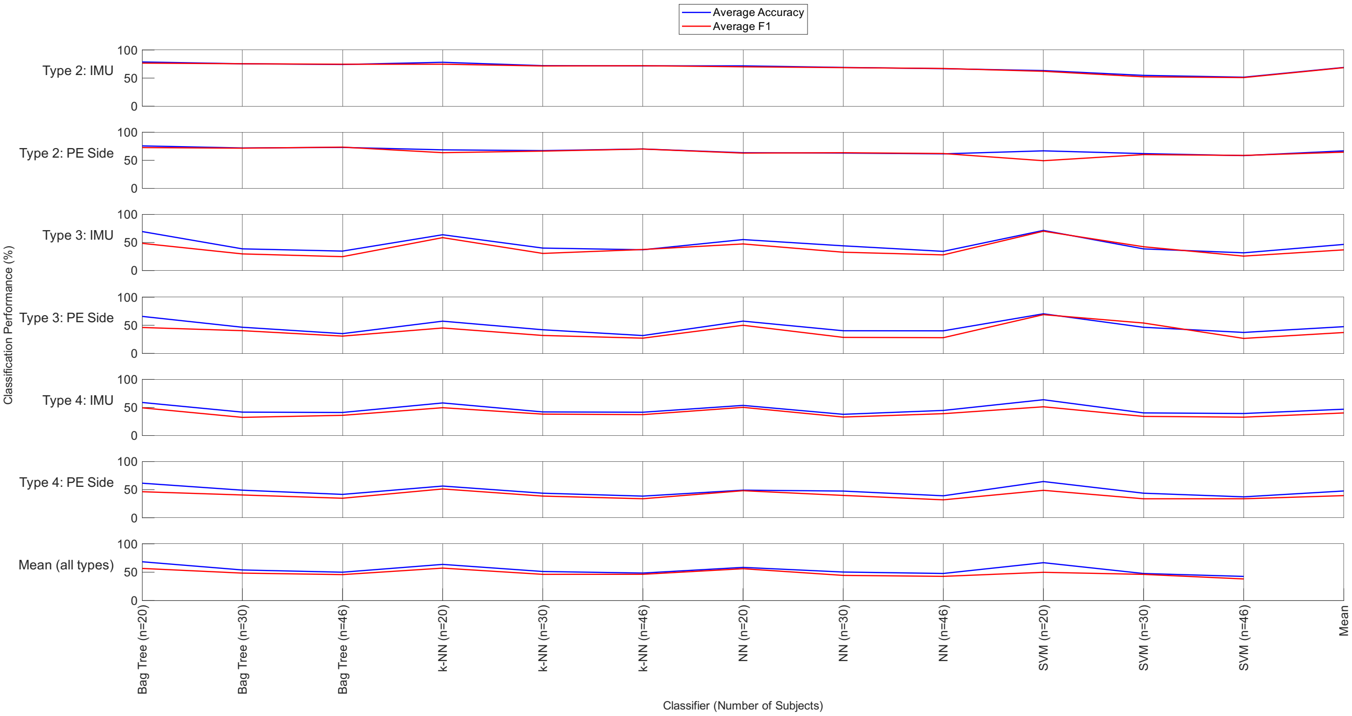

Figure 5: Development of the average accuracy and F_1 score for three cohorts.

Table 3: *Type 2* evaluated with and without known subjects.

n=46	**Bag Tree**	***k*-NN**
***Type 2*: IMU (46/46)**	74.19% (74.64%)	71.75% (72.02%)
***Type 2*: IMU (39/46)**	33.50% (29.65%)	36.11% (33.40%)
***Type 2*: PE Side (46/46)**	72.95% (73.44%)	70.15% (70.03%)
***Type 2*: PE Side (39/46)**	35.79% (32.18%)	30.65% (29.59%)
***Type 2*: PE Front (46/46)**	72.63% (73.09%)	71.78% (72.05%)
***Type 2*: PE Front (39/46)**	33.77% (28.98%)	29.35% (27.86%)

Cell format: { average accuracy (average F_1 score) }

Table 4: Mean results for all classifiers and evaluation types.

n=46	***Type 2***	***Type 3***	***Type 4***
Mean accuracy	68.00%	32.89%	38.97%
Mean F_1 score	68,26%	27.01%	33.47%

with subspace *k*-NN achieving the best overall results (78.37% accuracy and 78.34% F_1). For *Type 2*, classification performed best with the combined feature sets. The other three features sets performed about the same except for SVM which showed higher fluctuations. Accuracy and F_1 score are overall similar. For *Type 3*, all feature sets performed about the same, but compared to *Type 2*, the combined feature set performed slightly lower. F_1 scores are usually several percentage points lower than their corresponding accuracy values. NN achieved the highest accuracy result (40.13%), but the overall results are significantly lower than the results for *Type 2*. For *Type 4*, the feature set from IMU achieved the best results on average. The other three features sets performed slightly lower and are about the same. SVM achieved the highest accuracy result (44.78%). The overall results for *Type 4* are higher than the results for *Type 3*.

Table 4 presents the summed mean accuracy and F_1 score values for all classifiers and each evaluation type. *Type 2* has the highest mean results. *Type 4* has higher mean results than *Type 3*.

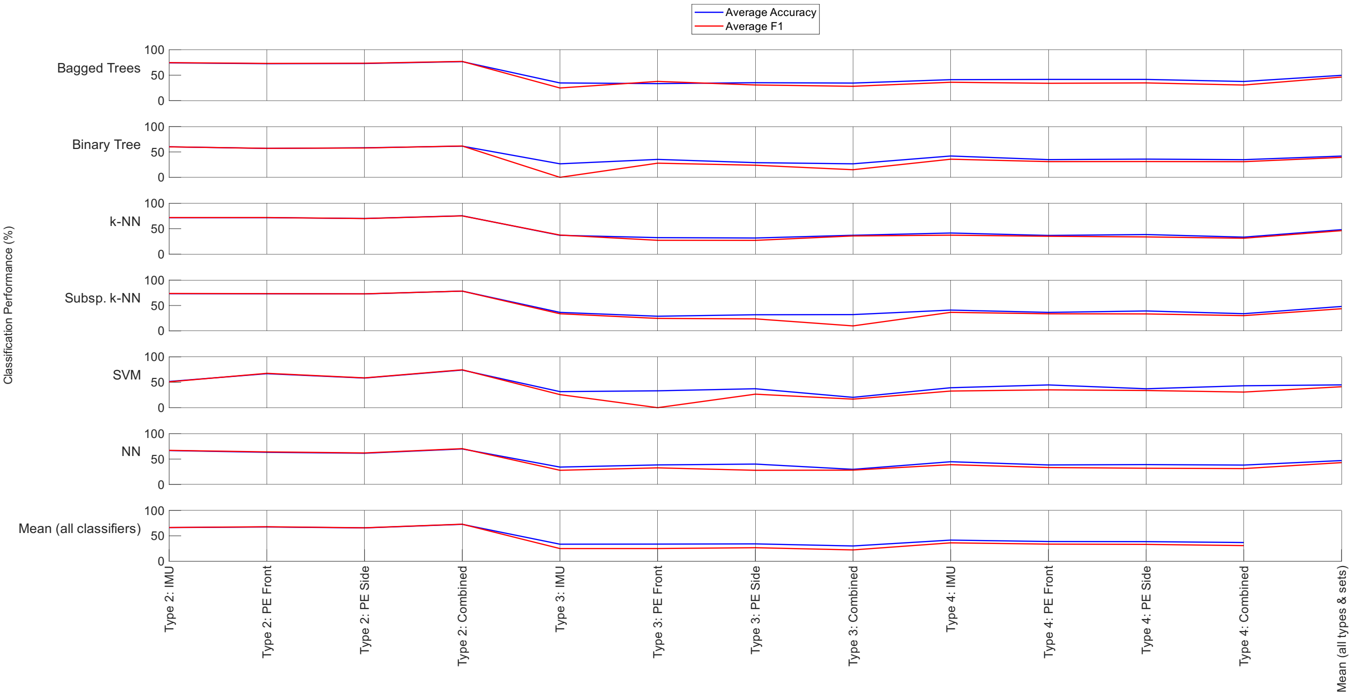

Figure 6: Average accuracy and F$_1$-score results for different evaluation types and data sets.

Table 5 demonstrates exemplarily for bagged trees how the results of a classifier depend on the particular test set (subjects), which is also true for all other classifiers. In particular, for *Type 3*, but also for *Type 4*, the results between best and worst classification results vary strongly for different test sets (subjects) compared to *Type 2*.

Table 5: Results for bagged trees with different test sets.

n=46	**Best**	**Worst**	**Average**
***Type 2*: IMU**	75.37% (76.17%)	72.92% (73.43%)	74.19% (74.64%)
***Type 2*: PE Front**	74.39% (74.52%)	71.13% (71.74%)	72.63% (73.09%)
***Type 2*: PE Side**	74.06% (74.52%)	72.27% (72.87%)	72.95% (73.44%)
***Type 2*: Comb.**	79.45% (79.99%)	73.57% (73.76%)	76.64% (77.03%)
***Type 3*: IMU**	64.41% (29.67%)	7.27% (19.88%)	34.69% (24.78%)
***Type 3*: PE Front**	52.78% (37.72%)	14.02% (37.72%)	33.38% (37.72%)
***Type 3*: PE Side**	50.85% (39.52%)	16.36% (21.90%)	35.16% (30.71%)
***Type 3*: Comb.**	61.02% (34.68%)	12.73% (21.55%)	34.45% (28.11%)
***Type 4*: IMU**	46.51% (40.90%)	31.63% (29.39%)	41.13% (36.03%)
***Type 4*: PE Front**	44.96% (42.85%)	36.04% (27.75%)	41.62% (34.01%)
***Type 4*: PE Side**	50.00% (36.41%)	35.24% (30.74%)	41.62% (34.76%)
***Type 4*: Comb.**	43.60% (36.72%)	31.33% (26.84%)	37.56% (30.57%)

Cell format: { accuracy (F$_1$ score) }

5 DISCUSSION

This section discusses the findings based on the results in Section 4. It is divided into four subsections: evaluation types, imbalanced data, oversampling, and limitations.

5.1 Evaluation Types

Type 2 classifiers achieved the highest average results, which is also consistent with the results in Wang and He [66]. A *Type 2* classifier is trained with some random data of each subject, so it knows each subject at least to some extent, which is also reflected in the test results (see Figure 5 and 6). As soon as such a trained classifier is confronted with data from unknown subjects, the results are significantly lower (see Table 3). The reason is that small data sets with a low number of subjects usually do not cover all possible variations (as also stated in [16, 18]). *Type 2* may only result in generalizable classifiers, if the target group and setting are so specific and homogeneous that a small data set already covers all possible variations – which can be difficult to prove. A further challenge is that motion signals may not only differ across subjects but also across the same subject in different trials [34]. Overall, *Type 2* classifiers will not generalize well when the small data set covers only a small portion of all possible variations; studies with machine learning and small data sets should narrow the target group and parameters in their study design to minimize the amount of all possible variation so that a small data set is able to cover most of that variation. This becomes evident when a trained classifier constantly has difficulty in classifying new subjects. Applying extensive feature engineering to improve accuracy would only amplify the specialization of the classifier to the particular small data set without improving its generality.

Type 4 achieved higher test results than *Type 3* (see Table 4). The reason for the difference in performance is probably that more diverse data from multiple subjects are used for the test sets in *Type 4* evaluation. The ratio between test and training set in *Type 3* depends on the total number of subjects and is often far below 5%. Table 5 shows that the range between best and worst results is higher for *Type 3* than for *Type 4*, demonstrating the instability and dependence of *Type 3* on the particular test set (subject). Cross-validation can be applied to *Type 3*, but the classification results are questionable due to the low test set ratio and the likely low generalizability of the trained classifier. We therefore do not recommend *Type 3* evaluations. Majority vote with multiple *Type 3* classifiers could be an option [4]. However, such an ensemble classifier would have been trained on both train and test data, which makes the evaluation of such a classifier (as with *Type 2*) questionable.

Regardless of the type of evaluation, the average results of all trained classifiers decrease as the number of subjects increases (see Figure 5). As discussed before, one possible explanation is that more subjects introduce more variance into the overall data set, which would require even more data to train a classifier. This also signifies that the trained classifiers do not generalize well at this point.

5.2 Imbalanced Data

In the context of fatigue research during physical exercise, data sets are prone to imbalanced classes because fewer data is usually collected for the fatigue state – subjects cannot perform an exercise in a fatigue state for an extended period of time. In most studies, subjects are required to perform the exercises to exhaustion which guarantees the fatigue state for each subject. An alternative is to perform exercises with a fixed time limit, which may result in some subjects not becoming fatigued due to differences in fitness levels. Fitness level and physical ability may also be the cause for a different number of recorded segments for each subject (see also Aguirre et al. [1]).

Some studies combine classes to improve classification results, e.g., [1, 22, 28, 36, 37, 40, 67]. Fewer classes lead to more data distributed across all classes, which is especially true for small data sets. As described by Zhang et al. [69, 70], the sample size, input dimension, and output classes need to be considered to avoid overfitting. Studies with small data sets should balance the number of samples with the number of classes accordingly. In general, the more classes, the more data is needed (see also Escobar-Linero et al. [22]). Regression can be a suitable alternative for detecting multiple fatigue levels (see also Shi et al. [61]). If a classifier predicts a class that is adjacent to the true class, the prediction is considered incorrect. Since fatigue is usually based on subjective scales, some tolerance for an incorrect prediction could be considered, e.g., if the actual RPE value is 15 and 13.8 is predicted on a scale of 6 to 20. With regression, more classes may be feasible.

A further aspect is how to reduce the number of classes. A commonly used method is to set a threshold and split the data into two groups for binary classification. Another common approach is binning where classes are grouped together into a smaller number of classes. Both approaches suffer from finding the right threshold(s) and distribution of classes (see Figure 4). Depending on how the new distribution of classes turns out, it impacts the results of the trained classifiers [62]. Some studies try to circumvent this, e.g., by using regression methods like mean absolute error as measure [65] (instead of accuracy or F_1 score) or by calculating the differences between successive RPE values [22]. The latter method does not work well when RPE values change linearly (as is the case in our study), resulting in a consistent delta around zero.

Another problem with small data can occur with cross-validation. Depending on how the data is partitioned, the distribution of classes within each partition can vary considerably. In extreme cases, certain classes may no longer occur in a partition. For example: a subject is selected as test set (*Type 3*) who has not reported RPE values above 12 on a scale of 6 to 20. Reducing the overall number of classes reduces the likelihood of this phenomenon.

5.3 Oversampling

A further consideration for studies on fatigue with small data sets is how to increase the data in the minority classes. For our study, two approaches were discussed. (1) Intra-subject oversampling creates new observations only from the same subject. Depending on the parameters, this approach creates small variations which may also be caused by sensor noise, and is therefore less artificial than the following approach. (2) Inter-subject oversampling creates new observations from multiple subjects of the same class. Depending on the oversampling parameters, the dimension of the feature vector, and the variation between subjects, the newly created observations may differ greatly from the original observations. This approach creates more diverse observations and may improve the generalizability of the classifier, but the new data may not occur in reality and can therefore affect the accuracy of the classifier in one way or another. Inter-subject oversampling is likely to generate data that increases the variance in the entire data set, which may paradoxically lead to the need of even more training data due to the higher variance. Oversampling can also result in an overweighting of data from certain subjects, e.g., if some subjects are only represented in the minority class(es) but not in other classes. Finding the right oversampling ratio is a further challenge but beyond the scope of this research.

5.4 Limitations

Our study has limitations that should be considered when interpreting our findings. The results show that classification performance strongly depends on several factors, including the number of subjects, the evaluation type, the classifier, class distribution, class imbalances, data homogeneity, and feature selection. To promote reproducibility and comparability across studies, it is crucial to provide detailed information about these factors.

One limitation of our study is that it was conducted in a laboratory setting, which may have influenced subjects' behavior and the collected data. As noted by Morris et al. [49], a laboratory environment that does not simulate a gym may produce different data than a real-world environment. Another limitation is the unbalanced ratio of male to female subjects (2:1), which could have affected the trained classifiers' ability to accurately detect fatigue levels in female participants. Furthermore, our target group is composed of non-athletes which probably results in a wider variance in the collected data compared to a group of (professional) athletes, as non-athletes may have more varied levels of fitness and fatigue tolerance. Additionally, our study focused only on the squat exercise; other exercises would have been equally valuable for the purpose of this study. Another limitation of our study is that machine learning was limited to statistical features.

Although we found that the feature sets from IMU and PE produced similar results (see Figure 5 and 6), suggesting that PE could be a viable alternative or complement to IMUs, further experiments are needed with different exercises to confirm these findings.

6 CONCLUSION

In our study, we have highlighted the challenges and potential pitfalls of utilizing machine learning in fatigue detection with small data sets, particularly due to subject variance, data partitioning, and evaluation methods. In our literature review, we found that related studies rarely report the impact of their methods on outcomes, indicating the need for more transparency when choosing particular research methods. Based on the results of our study with 46 subjects, we recommend researchers to avoid using *Type 2* (no unknown test subjects) and *Type 3* (leave one subject out) evaluations with subject-oriented small data sets, which may lead to models that lack generalizability, and instead utilize *Type 4* evaluations with leave n subjects out. Additionally, to maximize the value of small data sets, researchers should aim for a more homogeneous group of subjects and a balanced distribution of classes while minimizing potential variance. Although oversampling can help balance imbalanced data, it should be used with caution to avoid overly artificial data or overemphasis on certain subjects. In the future, we aim to identify the data saturation point for our study and hope that our findings will provide valuable guidance for other researchers in the field.

REFERENCES

[1] Andrés Aguirre, Maria J. Pinto, Carlos A. Cifuentes, Oscar Perdomo, Camilo A. R. Díaz, and Marcela Múnera. 2021. Machine Learning Approach for Fatigue Estimation in Sit-to-Stand Exercise. *Sensors* 21, 15 (2021). https://doi.org/10.3390/s21155006

[2] Wim Ament and Gijsbertus J. Verkerke. 2009. Exercise and Fatigue. *Sports Med.* 39, 5 (2009), 389–422. https://doi.org/10.2165/00007256-200939050-00005

[3] D. Anguita, A. Ghio, L. Oneto, X. Parra, and Jorge Luis Reyes-Ortiz. 2013. A Public Domain Dataset for Human Activity Recognition using Smartphones. In *Esann*.

[4] Amir Baghdadi, Fadel M. Megahed, Ehsan T. Esfahani, and Lora A. Cavuoto. 2018. A machine learning approach to detect changes in gait parameters following a fatiguing occupational task. *Ergonomics* 61, 8 (2018), 1116–1129. https://doi.org/10.1080/00140139.2018.1442936 arXiv:https://doi.org/10.1080/00140139.2018.1442936 PMID: 29452575.

[5] Konstantinos Balaskas and Kostas Siozios. 2021. Fatigue Detection Using Deep Long Short-Term Memory Autoencoders. In *2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. 1–4. https://doi.org/10.1109/MOCAST52088.2021.9493378

[6] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. *CoRR* abs/2006.10204 (2020). arXiv:2006.10204 https://arxiv.org/abs/2006.10204

[7] Daniel Berrar. 2019. Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology*, Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach (Eds.). Academic Press, Oxford, 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

[8] A. Bevilacqua, B. Huang, R. Argent, B. Caulfield, and T. Kechadi. 2018. Automatic classification of knee rehabilitation exercises using a single inertial sensor: A case study. In *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 21–24. https://doi.org/10.1109/bsn.2018.8329649

[9] Gürkan Bilgin, İ. Ethem Hindistan, Y. Gül Özkaya, Etem Köklükaya, Övünç Polat, and Ömer H. Çolak. 2015. Determination of Fatigue Following Maximal Loaded Treadmill Exercise by Using Wavelet Packet Transform Analysis and MLPNN from MMG-EMG Data Combinations. *Journal of Medical Systems* 39, 10 (Aug. 2015). https://doi.org/10.1007/s10916-015-0304-5

[10] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. 2021. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *CoRR* abs/2103.04922 (2021). arXiv:2103.04922 https://arxiv.org/abs/2103.04922

[11] G. A. Borg. 1982. Psychophysical bases of perceived exertion. *Medicine and science in sports and exercise* 14, 5 (1982), 377–381. https://pubmed.ncbi.nlm.nih.gov/7154893 7154893[pmid].

[12] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: Synthetic Minority Over-sampling Technique. *CoRR* abs/1106.1813 (2011). arXiv:1106.1813 http://arxiv.org/abs/1106.1813

[13] Louise Brennan, Antonio Bevilacqua, Tahar Kechadi, and Brian Caulfield. 2020. Segmentation of shoulder rehabilitation exercises for single and multiple inertial sensor systems. *Journal of Rehabilitation and Assistive Technologies Engineering* 7 (Jan. 2020), 205566832091537. https://doi.org/10.1177/2055668320915377

[14] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*. 3121–3124. https://doi.org/10.1109/ICPR.2010.764

[15] C. Buckley, M.A. O'Reilly, D. Whelan, A. Vallely Farrell, L. Clark, V. Longo, M.D. Gilchrist, and B. Caulfield. 2017. 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)*. 197–201. https://doi.org/10.1109/BSN.2017.7936040

[16] Achim Buerkle, Harveen Matharu, Ali Al-Yacoub, Niels Lohse, Thomas Bamber, and Pedro Ferreira. 2021. An adaptive human sensor framework for human–robot collaboration. *The International Journal of Advanced Manufacturing Technology* 119, 1-2 (Nov. 2021), 1233–1248. https://doi.org/10.1007/s00170-021-08299-2

[17] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. *ACM Comput. Surv.* 46, 3, Article 33 (Jan. 2014), 33 pages. https://doi.org/10.1145/2499621

[18] Christos Chalitsios, Thomas Nikodelis, Vasileios Konstantakos, and Iraklis Kollias. 2022. Sensitivity of movement features to fatigue during an exhaustive treadmill run. *European Journal of Sport Science* 22, 9 (2022), 1374–1382. https://doi.org/10.1080/17461391.2021.1955015 arXiv:https://doi.org/10.1080/17461391.2021.1955015 PMID: 34256682.

[19] Yeok Tatt Cheah, Ka Wing Frances Wan, and Joanne Yip. 2022. Prediction of Muscle Fatigue During Dynamic Exercises based on Surface Electromyography Signals Using Gaussian Classifier. In *Physical Ergonomics and Human Factors*. AHFE International. https://doi.org/10.54941/ahfe1002597

[20] Xilai Chen, Meiqin Liu, and Senlin Zhang. 2021. An LSTM-Attention-based Method to Muscle Fatigue Detection by Integrating Multi-Source sEMG Signals. In *2021 40th Chinese Control Conference (CCC)*. 8475–8480. https://doi.org/10.23919/CCC52363.2021.9549359

[21] Mohamed Elshafei, Diego Elias Costa, and Emad Shihab. 2021. On the Impact of Biceps Muscle Fatigue in Human Activity Recognition. *Sensors* 21, 4 (2021). https://doi.org/10.3390/s21041070

[22] Elena Escobar-Linero, Manuel Domínguez-Morales, and José Luis Sevillano. 2022. Worker's physical fatigue classification using neural networks. *Expert Systems with Applications* 198 (2022), 116784. https://doi.org/10.1016/j.eswa.2022.116784

[23] Julian J. Faraway and Nicole H. Augustin. 2018. When small data beats big data. *Statistics & Probability Letters* 136 (2018), 142–145. https://doi.org/10.1016/j.spl.2018.02.031 The role of Statistics in the era of big data.

[24] Alberto Fernández, Salvador García, Francisco Herrera, and Nitesh V. Chawla. 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary. *J. Artif. Int. Res.* 61, 1 (jan 2018), 863–905.

[25] Thomas Fontanari, Tiago Comassetto Fróes, and Mariana Recamonde-Mendoza. 2022. Cross-validation Strategies for Balanced and Imbalanced Datasets. In *Intelligent Systems*, João Carlos Xavier-Junior and Ricardo Araújo Rios (Eds.). Springer International Publishing, Cham, 626–640.

[26] Yuri Gordienko, Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Michail Novotarskiy, and Nikita Gordienko. 2018. Deep Learning for Fatigue Estimation on the Basis of Multimodal Human-Machine Interactions. https://doi.org/10.48550/ARXIV.1801.06048

[27] M. Guaitolini, L. Truppa, A. M. Sabatini, A. Mannini, and C. Castagna. 2020. Sport-induced fatigue detection in gait parameters using inertial sensors and support vector machines. In *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. 170–174. https://doi.org/10.1109/BioRob49111.2020.9224449

[28] Xiaole Guan, Yanfei Lin, Qun Wang, Zhiwen Liu, and Chengyi Liu. 2021. Sports fatigue detection based on deep learning. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 1–6. https://doi.org/10.1109/CISP-BMEI53629.2021.9624395

[29] Hao Guo, Li Ke, Qiang Du, and Song Guo. 2022. Muscle fatigue state classification based on blood flow bioimpedance. In *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 1–6. https://doi.org/10.1109/CISP-BMEI56279.2022.9980152

[30] Xiaonan Guo, Jian Liu, and Yingying Chen. 2020. When your wearables become your fitness mate. *Smart Health* 16 (May 2020), 100114. https://doi.org/10.1016/j.smhl.2020.100114

[31] Eric B. Hekler, Predrag Klasnja, Guillaume Chevance, Natalie M. Golaszewski, Dana Lewis, and Ida Sim. 2019. Why we need a small data paradigm. *BMC Medicine* 17, 1 (July 2019). https://doi.org/10.1186/s12916-019-1366-x

[32] Fauzani.N Jamaluddin, Siti A. Ahmad, Samsul Bahari Mohd Noor, Wan Zuha Wan Hassan, and E.F Shair. 2018. Performance of Different Threshold Estimation Methods on SEMG Wavelet De-noising in Prolonged Fatigue Identification. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. 293–296. https://doi.org/10.1109/IECBES.2018.8626599

[33] Yanran Jiang, Vincent Hernandez, Gentiane Venture, Dana Kulić, and Bernard K. Chen. 2021. A Data-Driven Approach to Predict Fatigue in Exercise Based on Motion Data from Wearable Sensors or Force Plate. *Sensors* 21, 4 (2021).

https://doi.org/10.3390/s21041499

[34] Yanran Jiang, Peter Malliaras, Bernard Chen, and Dana Kulic. [n. d.]. Real-time forecasting of exercise-induced fatigue from wearable sensors. 148 ([n. d.]), 105905. https://doi.org/10.1016/j.compbiomed.2022.105905

[35] P.A. Karthick, Diptasree Maitra Ghosh, and S. Ramakrishnan. 2018. Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and machine learning algorithms. *Computer Methods and Programs in Biomed.* 154 (2018), 45–56. https://doi.org/10.1016/j.cmpb.2017.10.024

[36] Swapnali Karvekar, Masoud Abdollahi, and Ehsan Rashedi. 2019. A Data-Driven Model to Identify Fatigue Level Based on the Motion Data from a Smartphone. *bioRxiv* (2019). https://doi.org/10.1101/796854 arXiv:https://www.biorxiv.org/content/early/2019/10/08/796854.full.pdf

[37] Swapnali Karvekar, Masoud Abdollahi, and Ehsan Rashedi. 2021. Smartphone-based human fatigue level detection using machine learning approaches. *Ergonomics* 64, 5 (2021), 600–612. https://doi.org/10.1080/00140139.2020.1858185 arXiv:https://doi.org/10.1080/00140139.2020.1858185 PMID: 33393439.

[38] Rob Kitchin and Tracey Lauriault. 2015. Small data in the era of big data. *GeoJournal* 80 (08 2015), 463–475. https://doi.org/10.1007/s10708-014-9601-7

[39] Yousef Kowsar, Masud Moshtaghi, Eduardo Velloso, Lars Kulik, and Christopher Leckie. 2016. Detecting Unseen Anomalies in Weight Training Exercises. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction* (Launceston, Tasmania, Australia) *(OzCHI '16)*. Association for Computing Machinery, New York, NY, USA, 517–526. https://doi.org/10.1145/3010915.3010941

[40] Jan Kuschan and Jörg Krüger. 2021. Fatigue recognition in overhead assembly based on a soft robotic exosuit for worker assistance. *CIRP Annals* 70, 1 (2021), 9–12. https://doi.org/10.1016/j.cirp.2021.04.034

[41] Arsalan Lambay, Ying Liu, Phillip Morgan, and Ze Ji. 2021. A Data-Driven Fatigue Prediction using Recurrent Neural Networks. In *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. 1–6. https://doi.org/10.1109/HORA52670.2021.9461377

[42] Dujuan Li and Caixia Chen. 2022. Research on exercise fatigue estimation method of Pilates rehabilitation based on ECG and sEMG feature fusion. *BMC Medical Informatics and Decision Making* 22, 1 (March 2022). https://doi.org/10.1186/s12911-022-01808-7

[43] J. F. Lin, M. Karg, and D. Kulić. 2016. Movement Primitive Segmentation for Human Motion Modeling: A Framework for Analysis. *IEEE Transactions on Human-Machine Systems* 46, 3 (2016), 325–339. https://doi.org/10.1109/thms.2015.2493536

[44] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[45] Luca Marotta, Bouke L. Scheltinga, Robbert van Middelaar, Wichor M. Bramer, Bert-Jan F. van Beijnum, Jasper Reenalda, and Jaap H. Buurke. 2022. Accelerometer-Based Identification of Fatigue in the Lower Limbs during Cyclical Physical Exercise: A Systematic Review. *Sensors* 22, 8 (2022). https://doi.org/10.3390/s22083008

[46] Neusa R. Adão Martins, Simon Annaheim, Christina M. Spengler, and René M. Rossi. 2021. Fatigue Monitoring Through Wearables: A State-of-the-Art Review. *Frontiers in Physiology* 12 (Dec. 2021). https://doi.org/10.3389/fphys.2021.790292

[47] V.F. Milanez, M.C. Spiguel Lima, C.A. Gobatto, L.A. Perandini, F.Y. Nakamura, and L.F.P. Ribeiro. 2011. Correlates of session-rate of perceived exertion (RPE) in a karate training session. *Science & Sports* 26, 1 (2011), 38–43. https://doi.org/10.1016/j.scispo.2010.03.009

[48] Ahmad Moniri, Dan Terracina, Jesus Rodriguez-Manzano, Paul H. Strutton, and Pantelis Georgiou. 2021. Real-Time Forecasting of sEMG Features for Trunk Muscle Fatigue Using Machine Learning. *IEEE Transactions on Biomedical Engineering* 68, 2 (2021), 718–727. https://doi.org/10.1109/TBME.2020.3012783

[49] Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Acm. https://doi.org/10.1145/2556288.2557116

[50] Farnad Nasirzadeh, Mostafa Mir, Sadiq Hussain, Mohammad Tayarani Darbandy, Abbas Khosravi, Saeid Nahavandi, and Brad Aisbett. 2020. Physical Fatigue Detection Using Entropy Analysis of Heart Rate Signals. *Sustainability* 12, 7 (2020). https://doi.org/10.3390/su12072714

[51] 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 *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 606–615. https://doi.org/10.1145/3219819.3219864

[52] Michalis Papakostas, Varun Kanal, Maher Abujelala, Konstantinos Tsiakas, and Fillia Makedon. [n. d.]. Physical fatigue detection through EMG wearables and subjective user reports: a machine learning approach towards adaptive rehabilitation. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes Greece, 2019-06-05). ACM, 475–481. https://doi.org/10.1145/3316782.3322772

[53] Kun Peng. 2022. Training Fatigue and Recovery of Throwing Athletes Based on the Comprehensive Environmental Test of the Field. *Wireless Communications and Mobile Computing* 2022 (Aug. 2022), 1–10. https://doi.org/10.1155/2022/7993666

[54] Ross O. Phillips. 2015. A review of definitions of fatigue – And a step towards a whole definition. *Transportation Research Part F: Traffic Psychology and Behaviour* 29 (2015), 48–56. https://doi.org/10.1016/j.trf.2015.01.003

[55] Neelam Rout, Debahuti Mishra, and Manas Kumar Mallick. 2018. Handling Imbalanced Data: A Survey. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, M. Sreenivasa Reddy, K. Viswanath, and Shiva Prasad K.M. (Eds.). Springer Singapore, Singapore, 431–443.

[56] ZahraAlizadeh Sani, MohammadTayarani Darbandy, Mozhdeh Rostamnezhad, Sadiq Hussain, Abbas Khosravi, and Saeid Nahavandi. 2020. A new approach to detect the physical fatigue utilizing heart rate signals. *Research in Cardiovascular Medicine* 9, 1 (2020), 23. https://doi.org/10.4103/rcm.rcm_8_20

[57] Yutaka Sasaki. 2007. The truth of the F-measure. *Teach Tutor Mater* (01 2007).

[58] Gerd Schmitz. 2020. Moderators of Perceived Effort in Adolescent Rowers During a Graded Exercise Test. *International Journal of Environmental Research and Public Health* 17, 21 (Nov. 2020), 8063. https://doi.org/10.3390/ijerph17218063

[59] Zahra Sedighi Maman, Ying-Ju Chen, Amir Baghdadi, Seamus Lombardo, Lora A. Cavuoto, and Fadel M. Megahed. 2020. A data analytic framework for physical fatigue management using wearable sensors. *Expert Systems with Applications* 155 (2020), 113405. https://doi.org/10.1016/j.eswa.2020.113405

[60] M. Seiffert, F. Holstein, R. Schlosser, and J. Schiller. 2017. Next Generation Cooperative Wearables: Generalized Activity Assessment Computed Fully Distributed Within a Wireless Body Area Network. *IEEE Access* 5 (2017), 16793–16807. https://doi.org/10.1109/access.2017.2749005

[61] Song Shi, Ziping Cao, Hengheng Li, Chengming Du, Qiang Wu, and Yahui Li. 2022. Recognition System of Human Fatigue State Based on Hip Gait Information in Gait Patterns. *Electronics* 11, 21 (2022). https://doi.org/10.3390/electronics11213514

[62] Vimalraj S Spelmen and R Porkodi. 2018. A Review on Handling Imbalanced Data. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. 1–11. https://doi.org/10.1109/ICCTCT.2018.8551020

[63] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. 2009. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence* 23, 04 (June 2009), 687–719. https://doi.org/10.1142/s0218001409007326

[64] Mamello Thinyane. 2017. Investigating an Architectural Framework for Small Data Platforms. In *Data for societal challenges-17th European Conference on Digital Government (ECDG 2017)*. 220–227.

[65] Andreas Triantafyllopoulos, Sandra Ottl, Alexander Gebhard, Esther Rituerto-González, Mirko Jaumann, Steffen Hüttner, Valerie Dieter, Patrick Schneeweiß, Inga Krauß, Maurice Gerczuk, Shahin Amiriparian, and Björn W. Schuller. 2022. Fatigue Prediction in Outdoor Running Conditions using Audio Data. https://doi.org/10.48550/ARXIV.2205.04343

[66] Bin Wang and Dongzhi He. 2021. Prediction method of running fatigue based on depth image. In *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Vol. 4. 271–275. https://doi.org/10.1109/IMCEC51613.2021.9482204

[67] Guodong Wang, Xiaokun Mao, Qiuxia Zhang, and Aming Lu. [n. d.]. Fatigue Detection in Running with Inertial Measurement Unit and Machine Learning. In *2022 10th International Conference on Bioinformatics and Computational Biology (ICBCB)* (Hangzhou, China, 2022-05-13). IEEE, 85–90. https://doi.org/10.1109/ICBCB55259.2022.9802471

[68] Zhongwan Yang and Huijie Ren. 2019. Feature Extraction and Simulation of EEG Signals During Exercise-Induced Fatigue. *IEEE Access* 7 (2019), 46389–46398. https://doi.org/10.1109/ACCESS.2019.2909035

[69] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. https://doi.org/10.48550/ARXIV.1611.03530

[70] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding Deep Learning (Still) Requires Rethinking Generalization. *Commun. ACM* 64, 3 (feb 2021), 107–115. https://doi.org/10.1145/3446776

[71] Fan Zhang and Feng Wang. 2020. Exercise Fatigue Detection Algorithm Based on Video Image Information Extraction. 8 (2020), 199696–199709. https://doi.org/10.1109/ACCESS.2020.3023648

[72] Jian Zhang, Thurmon E. Lockhart, and Rahul Soangra. 2013. Classifying Lower Extremity Muscle Fatigue During Walking Using Machine Learning and Inertial Sensors. *Annals of Biomedical Engineering* 42, 3 (Oct. 2013), 600–612. https://doi.org/10.1007/s10439-013-0917-0

[73] Haiyan Zhu, Yuelong Ji, Baiyang Wang, and Yuyun Kang. 2022. Exercise fatigue diagnosis method based on short-time Fourier transform and convolutional neural network. *Frontiers in Physiology* 13 (Aug. 2022), 12 pages. https://doi.org/10.3389/fphys.2022.965974

[74] Min Zhu, Jing Xia, Xiaoqing Jin, Molei Yan, Guolong Cai, Jing Yan, and Gangmin Ning. 2018. Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data. *IEEE Access* 6 (2018), 4641–4652. https://doi.org/10.1109/ACCESS.2018.2789428