Digital
Biomarkers

Digit Biomark 2020;4(suppl 1):59–72

DOI: 10.1159/000512166
Received: July 24, 2020
Accepted: October 6, 2020
Published online: November 26, 2020

Emerging Applications

Assessment of Fatigue Using Wearable Sensors: A Pilot Study

Hongyu Luo[a] Pierre-Alexandre Lee[a] Ieuan Clay[b] Martin Jaggi[c]
Valeria De Luca[a]

[a]Novartis Institutes for Biomedical Research, Basel, Switzerland; [b]Evidation Health Inc., San Mateo, CA, USA; [c]Machine Learning and Optimization Laboratory, EPFL, Lausanne, Switzerland

Keywords

Wearable sensors · Fatigue · Machine learning · Deep learning · Digital measurements

Abstract

Background: Fatigue is a broad, multifactorial concept encompassing feelings of reduced physical and mental energy levels. Fatigue strongly impacts patient health-related quality of life across a huge range of conditions, yet, to date, tools available to understand fatigue are severely limited. ***Methods:*** After using a recurrent neural network-based algorithm to impute missing time series data form a multisensor wearable device, we compared supervised and unsupervised machine learning approaches to gain insights on the relationship between self-reported non-pathological fatigue and multimodal sensor data. ***Results:*** A total of 27 healthy subjects and 405 recording days were analyzed. Recorded data included continuous multimodal wearable sensor time series on physical activity, vital signs, and other physiological parameters, and daily questionnaires on fatigue. The best results were obtained when using the causal convolutional neural network model for unsupervised representation learning of multivariate sensor data, and random forest as a classifier trained on subject-reported physical fatigue labels (weighted precision of 0.70 ± 0.03 and recall of 0.73 ± 0.03). When using manually engineered features on sensor data to train our random forest (weighted precision of 0.70 ± 0.05 and recall of 0.72 ± 0.01), both physical activity (energy expenditure, activity counts, and steps) and vital signs (heart rate, heart rate variability, and respiratory rate) were important parameters to measure. Furthermore, vital signs contributed the most as top features for predicting mental fatigue compared to physical ones. These results support the idea that fatigue is a highly multimodal concept. Analysis of clusters from sensor data highlighted a digital phenotype indicating the presence of fatigue (95% of observations) characterized by

Valeria De Luca
Novartis Institutes for Biomedical Research
Fabrikstrasse 2, Novartis Campus
CH–4056 Basel (Switzerland)
valeria.de_luca@novartis.com

a high intensity of physical activity. Mental fatigue followed similar trends but was less predictable. Potential future directions could focus on anomaly detection assuming longer individual monitoring periods. ***Conclusion:*** Taken together, these results are the first demonstration that multimodal digital data can be used to inform, quantify, and augment subjectively captured non-pathological fatigue measures.

Introduction

Improving patients' health and health-related quality of life (HRQoL) [1] is a primary aim for all therapies: perceived benefits of treatment are the basis for both their registration and marketing. Patient-reported outcomes (PROs) are included in most clinical trials as instruments to assess whether patients perceive improvements in their disease symptoms. PROs are typically collected as questionnaires, usually presented as multiple choice or a number from a predefined range. Despite their widespread use, PROs have several well-established shortcomings, including their sparsity and lack of reproducibility, as well as practical challenges such as low compliance [2]. Efforts to address these challenges focus on improving the delivery of PROs by using electronic diaries (although their benefit is still debated [3]) and statistical models to better control for biases [4].

An emerging, alternative avenue for improving our ability to assess HRQoL is to use objective data, for example collected via sensors, to create alternate measures of relevant symptoms [5, 6]. Research in this area is most advanced in creating digital measures of function, for example monitoring changes in mobility [7] or cognition [8, 9]. Similar methods are now also being applied to establish proof-of-principle to create objective, orthogonal measures to augment PROs which measure perceived changes in symptom relief and aspects of HRQoL, including mobility [10], pain [11, 12], stress [13, 14], and mood [15]. Most work so far has focused on domains where feature sets are derived from a single mode of sensor data, for example inertial sensors for mobility or galvanic skin response for stress; however, most HRQoL domains are extremely multifactorial and therefore require multimodal data.

An underexplored, multifactorial HRQoL factor is fatigue. The elements driving individual perceptions of fatigue are highly multifaceted and manifest as lower physical performance, somnolence, or reduced attention. Fatigue is experienced by diseased and healthy individuals and thus can be classified as pathological fatigue and non-pathological fatigue [16, 17]. Mental and physical fatigue exist in both pathological and non-pathological fatigue [16]. Pathological fatigue is a major symptom in cancer [18] and several neurological [17, 19, 20], autoimmune-related diseases and immunodeficiency disorders, such as multiple sclerosis and lupus [21]. Non-pathological fatigue is experienced in healthy subjects under various circumstances, such as working stress [22] and jet lag [23, 24]. Few recent studies have pioneered the use of wearable sensors to assess physical fatigue in healthy cohorts under experimental settings, for example inertial measurement units and heart rate monitors in construction workers [25–27] and runners [28].

In this study, we explore the relationship between self-reported, non-pathological physical and mental fatigue and behavioral and physiological multivariate time-series data acquired from a multisensor wearable device in healthy subjects in free living. Machine learning and deep learning approaches have been developed for time series classification, such as residual networks and multichannel deep convolutional neural networks (CNNs) [29]. Unsupervised methods have been proposed to generate general-purpose features from raw sensor data, which can be used together with more established statistical summaries and spectral features as input, for example for classification. Examples of recent approaches are causal CNNs (cCNNs) [30] and denoising autoencoders [31].

We compare several methods to predict non-pathological fatigue scores from physical activity and vital sign parameters, and further explore behavior and physiology patterns from wearable data and their connection to self-reported fatigue via clustering. The main contributions of this work are the development of a machine learning-based analysis framework to connect multimodal wearable sensor data, continuously acquired in real-world settings, to PROs, and the assessment of the most relevant digital measures to augment sporadically deployed, traditional PRO instruments. While the latter are collected mostly during clinical visits with healthcare professionals, sensor data can be used to deliver more frequent, reliable, and objective insights to assess several behavioral factors of HRQoL in clinical trials and clinical practice.

Materials and Methods

Data Acquisition

Data from 28 healthy individuals (26–55 years of age, average age 42 years, 41/51% female/male), of which 17 enrolled up to 2 days after returning from long-haul flights with 3–7 time zone differences and hence were recovering from jet lag, from 1 to 219 consecutive days (mean 35, median 9, total 973 days) were collected. Objective data were collected using a multisensor wearable device, Everion (Biovotion AG, Switzerland; https://www.biovotion.com/everion/), in conjunction with a mobile app, SymTrack (Gastric GmbH, Switzerland), to deliver a daily fatigue questionnaire.

Volunteers were asked to continuously wear the Everion device around their non-dominant arm over a 1-week period. The device combines a 3-axis accelerometer, barometer, galvanic skin response electrode, and temperature and photo sensors. We tracked a total of 12 parameters at 1-Hz temporal resolution on physical activity and physiology (see Table 1).

Volunteers were instructed to complete a 4-item daily questionnaire in the evening to capture their subjective assessment of fatigue, adapted from the Fatigue Assessment Scale [32] and Visual Analogue Scale to evaluate fatigue severity [33]:

1. Physical fatigue score (*PhF*). Question: Physically, today how often did you feel exhausted? Possible answers: never; sometimes; regularly; often; always.
2. Mental fatigue score (*MF*). Question: Mentally, today how often did you feel exhausted? Same answers as above.
3. Visual analogue scale score (*VAS*). Question: Describe fatigue on a scale of 1–10, where 1 means you don't feel tired at all and 10 means the worst tiredness you can imagine.
4. Indicator of relative perception (*RelP*). Question: Are you feeling better, worse, or the same as yesterday?

Further details and subject-level statistics on the collected data are described in the online supplementary materials (see www.karger.com/doi/10.1159/000512166 for all online suppl. material).

Data Pre-Processing

Parameters from the multisensor wearable device were downsampled to 1-min temporal resolution following the manufacturer's guidelines. For each subject and parameter, we excluded days where more than 80% of the 1,440 samples were missing to ensure an acceptable performance of downstream analysis [34]. Missing samples were due to subjects not wearing the device (e.g., during charging) or low-quality segments (e.g., loss of skin contact). This filtering step led to a total of 27 subjects and 530 days of data. Finally, we

Table 1. List of parameters of the multisensor wearable device (Biovotion Everion), aggregation approach for downsampling, and computed daily summary features

Sensor	Parameter	Description	Unit	Downsampling	Daily features
Accelerometer	ActivityClass	Categorical parameter. Type of physical activity: 0 = undefined, 1 = resting, 9 = other, 10 = biking, 11 = running, 12 = walking	–	Mode	Count per category; One-hot encoding
	ActivityCounts	The activity value indicates the intensity of motion (activity)	–	Sum	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD, Sum
	Steps	Number of steps	–	Sum	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD, Sum
	EnergyExpenditure	Amount of energy a person uses to complete all regular bodily functions, measured in calories	Calories/s	Sum	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD, Sum
Photoplethysmography	HR	Heart Rate	bpm	Quality-weighted average	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD
	HRV	Heart rate variability. Indicates the beat to beat variations.	ms	Quality-weighted average	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD
	RESP	Respiration rate. Number of breaths a person takes per minute	bpm	Quality-weighted average	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD
	BloodPerfusion	Blood perfusion can be measured as the percentage change in blood volume in local tissue resulting from a heartbeat	–	Median	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD
	BloodPulseWave	Blood is ejected generating a pulse wave when the heart contracts	–	Median	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD
Temperature	SkinTemperature	Skin temperature	°C	Median	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD
Galvanic Skin Response	GalvanicSkinResponse	Describes changes in the electrical conductivity of the skin. It is a measure of emotional arousal	kOhm	Mean	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD
Barometer	Barometer	Barometric pressure measures changes of altitude	mbar	Median	Mean, SD, Median, Max., Min., Skewness, Kurtosis, 5th and 95th percentile, FFT, PSD

HR, heart rate; HRV, heart rate variability; RESP, respiratory rate; SD, standard deviation; Max., maximum; Min., minimum; FFT, fast Fourier transform (amplitude and frequency of 2nd to 5th peaks); PSD, power spectral density (amplitude and frequency of 2nd to 5th peaks).

imputed missing data gaps using the state-of-the-art unidirectional uncorrelated recurrent imputation model from Cao et al. [34]. Details and evaluation of this model are given in the online supplementary materials.

Sensor Features

After pre-processing, $F = 254$ statistical summaries and spectral features were computed per observation (day) from the 12 parameters and are listed in Table 1. These features include the encoding of *RelP* ∈ {*worse*; *same*; *better*} to {−1;0;1} and the time zone difference (TZ_{diff}). On the day a subject changes their time zone, TZ_{diff} is equal to the absolute difference in hours between the time zones of the current and previous location; for the following days, TZ_{diff} will reduce 1 h each day until reaching 0.

The multidimensionality of the wearable device's time series makes the selection of relevant features a difficult task. Representation learning reduces data to a lower dimensional representation that could be further used as input for classification. We used the unsupervised cCNN model proposed in Franceschi et al. [30] to learn a general-purpose representation of multivariate sensor parameters. The model combines an encoder-only architecture based on causal dilated convolutions with a triplet loss inspired by word2vec [35] ensuring that similar time series get similar representations. Model and training details are provided in the online supplementary materials. Input time series of 1-min resolution were modified by one-hot encoding the categorical ActivityClass parameter. We empirically selected output encodings of length 40. *RelP* and TZ_{diff} were added to the feature vector of length $F_{cCNN} = 42$ for each observation.

Fatigue Labels

For simplicity and to reduce intra- and inter-rater variability, which is a known limitation of PROs [36], fatigue scores *PhF* and *MF* were intuitively converted into binary labels y_{PhF} and y_{MF}, respectively:

$$y = \begin{cases} 0 & \text{for } F = never \\ 1 & \text{for } F \in \{\text{sometimes; regularly; often; always}\} \end{cases}$$

In addition, binary labels y_{VAS} were computed on the *VAS* fatigue score [37, 38]:

$$y_{VAS} = \begin{cases} 0 & \text{for } VAS \in \{1; \ldots ; 4\} \\ 1 & \text{for } VAS \in \{5; \ldots ; 10\} \end{cases}$$

Fatigue labels are assigned to the same calendar day as sensor data. We selected only days when PROs were completed, resulting in a total of $D = 405$ days from 27 subjects.

Classification

To demonstrate whether subjective non-pathological mental and physical fatigue scores could be linked to physical activity, vital signs, and physiological parameters from wearable sensors, we compared random forest (RF) trained on the 2 main feature sets $X \in \mathrm{R}^{D \times F}$ (statistical summaries vs. cCNN encodings) to predict daily y_{PhF} and y_{MF}. Evaluation, hyperparameters, training, and cross-validation details for these models are described in the online supplementary materials. We computed accuracy, weighted precision, recall, and F1 to evaluate classification results, the latter being suitable to minimize skew-biased results due to imbalance labels [39].

Anomaly Detection

Prediction of fatigue episodes can be formulated as an outlier or anomaly detection problem given the low availability of high fatigue labels in the dataset. We used local outlier

DOI: 10.1159/000512166

factor (LOF) [40] to learn the local density distributions of daily observations and classify observations in low-density regions as outliers. We split the dataset into 2 parts: normal (with *PhF* or *MF* ∈ {*never*; *sometimes*; *regularly*}) and anomaly observations ({*often*; *always*}). LOF was trained on 90% of the normal observations and evaluated on the anomaly observations set and the remaining 10% of normal ones (test set) using the average error between anomaly and test sets. The classification error is defined as the percentage of observations classified as outliers in the test set and as inliers in the anomaly set. Implementation details are described in the online supplementary materials.

Clustering

To further investigate the relationship between objective behavioral data and subjective non-pathological fatigue scores and extract potential digital phenotypes of fatigue, we applied Ward's agglomerative hierarchical clustering [41] with Euclidean distance on a subset of $\boldsymbol{X}$, $X_c \in \mathrm{R}^{D \times Fc}$ with $F_c = 44$ from the mean, median, minimum and maximum of the 11 numerical parameters (excluding ActivityClass).

Data Availability

Data and code will be made publicly available at https://zenodo.org/ [42].

Results

We explored the relationship between subjectively reported non-pathological physical and mental fatigue and objective physiological and behavioral parameters from a number of perspectives: direct correlation, direct classification of PRO scores, anomaly detection (classification of outlying periods of data corresponding to increased fatigue reporting), and unsupervised clustering.

Dataset Statistics and Direct Correlation

Figure 1 shows value distributions of the 12 sensor parameters in their original units and 3 fatigue scores for all subjects. Subject-specific statistics are provided in the online supplementary materials. The only parameters with moderate correlation (absolute Spearman's rank correlation coefficient $0.5 < \rho < 0.7$ [43]) were HR with ActivityCounts and RESP. We also computed the correlation between *PhF*, *MF*, and *VAS* and the daily average of the numerical sensor parameters. All features have low-to-negligible correlation with fatigue scores ($\rho \leq 0.27$).

Classification of Fatigue Scores

Table 2 summarizes classification results on the validation set for predicting y_{PhF} and y_{MF}. Overall, better results were obtained when predicting physical fatigue compared to mental fatigue. While accuracy results between cCNN and statistical summary features were not statistically significantly different (McNemar's test [44] p value = 0.57/0.50 for physical/mental fatigue), cCNN + RF reached the highest average weighted F1 score of 0.71 ± 0.04. The parameters contributing the most to RF predictions were EnergyExpenditure, HRV, ActivityCounts, RESP, HR, and Steps. Power spectral density amplitudes were the top predictive features calculated from these parameters. The first 10% of features from RF are listed in the online supplementary materials.

DOI: 10.1159/000512166

www.karger.com/dib

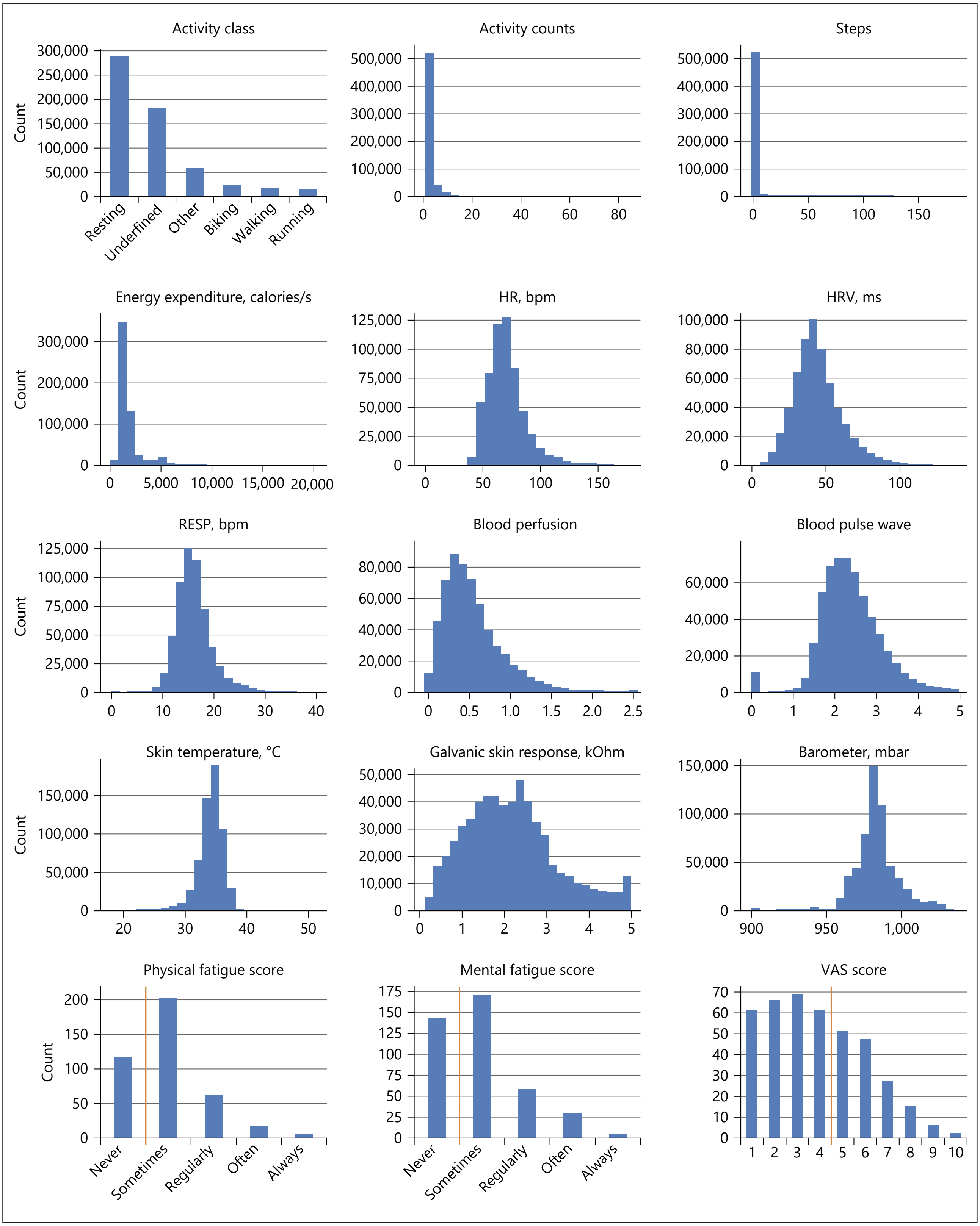

Fig. 1. Distribution of the 12 sensor parameters (1-min temporal resolution, top 4 rows) and physical, mental, and VAS fatigue scores (bottom row). Vertical orange lines indicate the thresholds used to compute fatigue labels.

Table 2. Prediction results of binary labels of physical and mental fatigue

Label	Model	Accuracy, %	McNemar's *p* value	Weighted precision	Weighted recall	Weighted F1 score
Physical fatigue	RF **cCNN + RF**	**71.85±1.44** 71.40±3.92	0.57	0.70±0.05 **0.70±0.03**	0.72±0.01 **0.73±0.03**	0.65±0.02 **0.71±0.04**
Mental fatigue	RF **cCNN + RF**	64.69±3.37 **66.20±4.58**	0.50	0.63±0.04 **0.65±0.05**	0.65±0.03 **0.66±0.05**	0.62±0.02 **0.65±0.05**

Evaluation metrics are summarized as the cross-validation mean ± SD. The best results are highlighted in bold font.

Table 3. Novelty detection results

Fatigue score	Experiment	Test set error (false positive)	Novelty set error (false negative)	Mean error
Physical fatigue	Daily summary	0.49	0.48	0.48
	Daily simple summary	0.25	0.65	0.45
	6-h simple summary	**0.36**	**0.43**	**0.40**
Mental fatigue	Daily summary	0.41	0.51	0.46
	Daily simple summary	0.35	0.57	0.46
	6-h simple summary	**0.51**	**0.40**	**0.46**

The best results are highlighted in bold font.

Anomaly Detection

Prediction of periods of non-pathological fatigue can also be modeled as an anomaly detection problem, identifying the relatively rare high fatigue labels as outliers in the dataset. Three experiments were performed using as input data: (i) X, (ii) X_c, and (iii) mean, median, minimum, and maximum computed over non-overlapping 6-h time windows in each calendar day, and concatenated for each day, $X_{6h} \in \mathrm{R}^{D \times F6h}$, with $F_{6h} = 176$. Overall, LOF achieved lower false positive rates than false negative rates. High mental fatigue observations were more difficult to detect than physical ones. The lowest mean error (40%) was achieved while trying to detect high *PhF* scores from the shorter, 6-h summaries (Table 3).

Cluster Analysis

We also used unsupervised methods in order to uncover general patterns in the data and explore their relation to fatigue labels. Figure 2 shows the output of hierarchical clustering computed on X_c, where each row corresponds to one observation and each column to the Z-score normalized values of the 44 features in X_c. For each observation, the corresponding fatigue labels and demographics are also displayed. We observed 3 clustered feature sets and 3 groups of 173, 114, and 118 observations each, details of which are provided in the online supplementary materials. Figure 3 shows the distribution of age, gender, and fatigue score in the 3 clusters. Group 3 is mainly composed of male subjects (89%) from the youngest age range (80% of age range 26–35 years), mostly includes days where fatigue was reported (95% of observations with $y_{PhF} = 1$ and 75% with $y_{MF} = 1$), and shows significantly higher (t test with Bonferroni correction, $p \leq 1.00e{-}04$) physical activity readouts, maximum HR, and lower HRV than the other groups. We also observed that group 1 had similar percentages of fatigue and non-fatigue observations but significantly higher mean HR, RESP, BloodPerfusion, BloodPulseWave, and GalvanicSkinResponse than the other 2 groups, suggesting that subjects in this group were undergoing stressful events during the day [45].

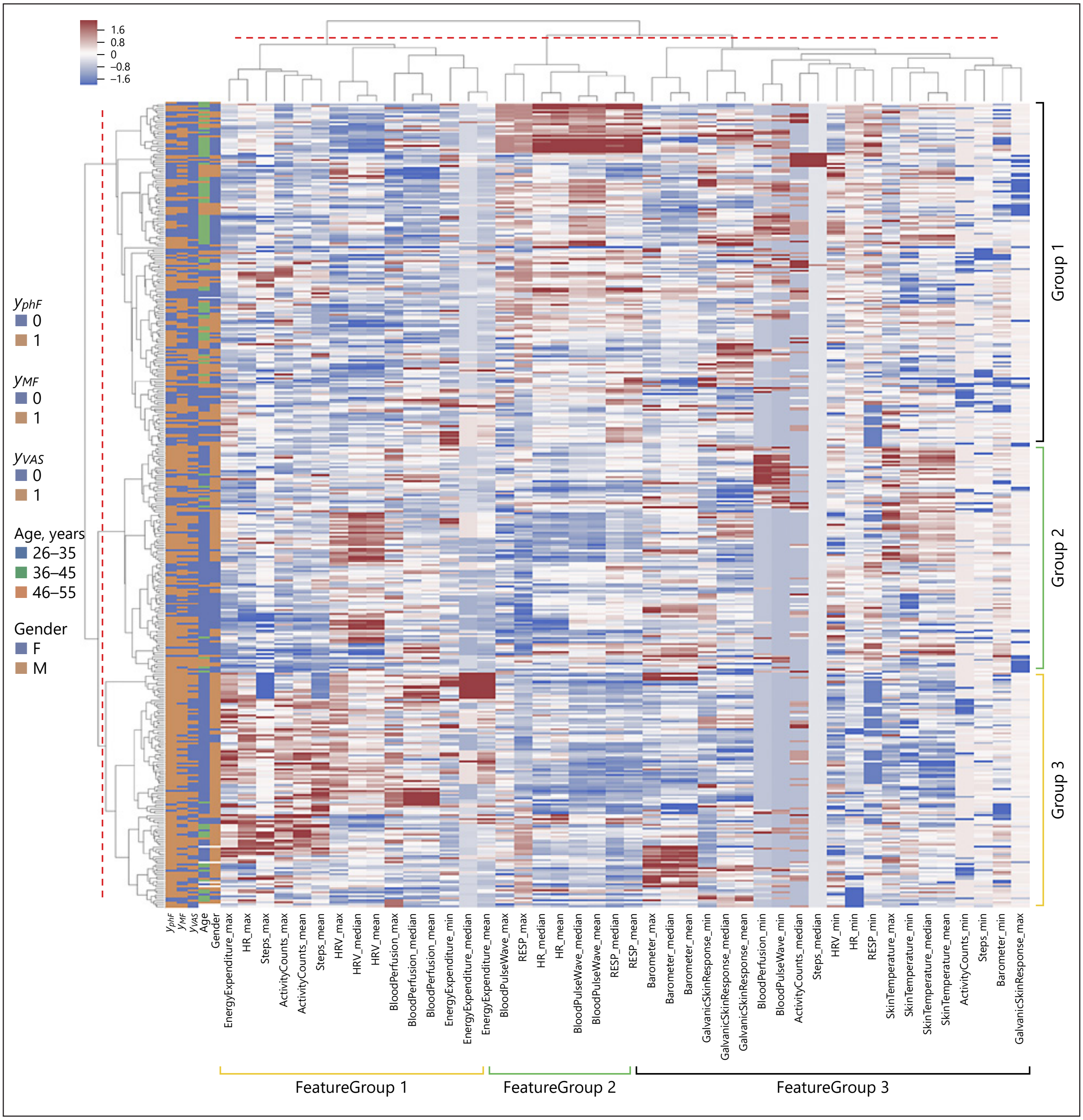

Fig. 2. Hierarchical clustering overview with dendrograms and heatmap of ranked features. Resulting clusters (Group 1, 2, and 3) and input feature sets (FeatureGroup 1, 2, and 3) at dendrograms height = 40 and 50, respectively, are highlighted.

Discussion

Classification Performance

Compared to classical benchmark models, such as support vector machine (see online suppl. materials), RF outperformed in weighted F1 score in all experiments due to its ability to deal with highly dimensional, unbalanced datasets. By extracting the top predictive statis-

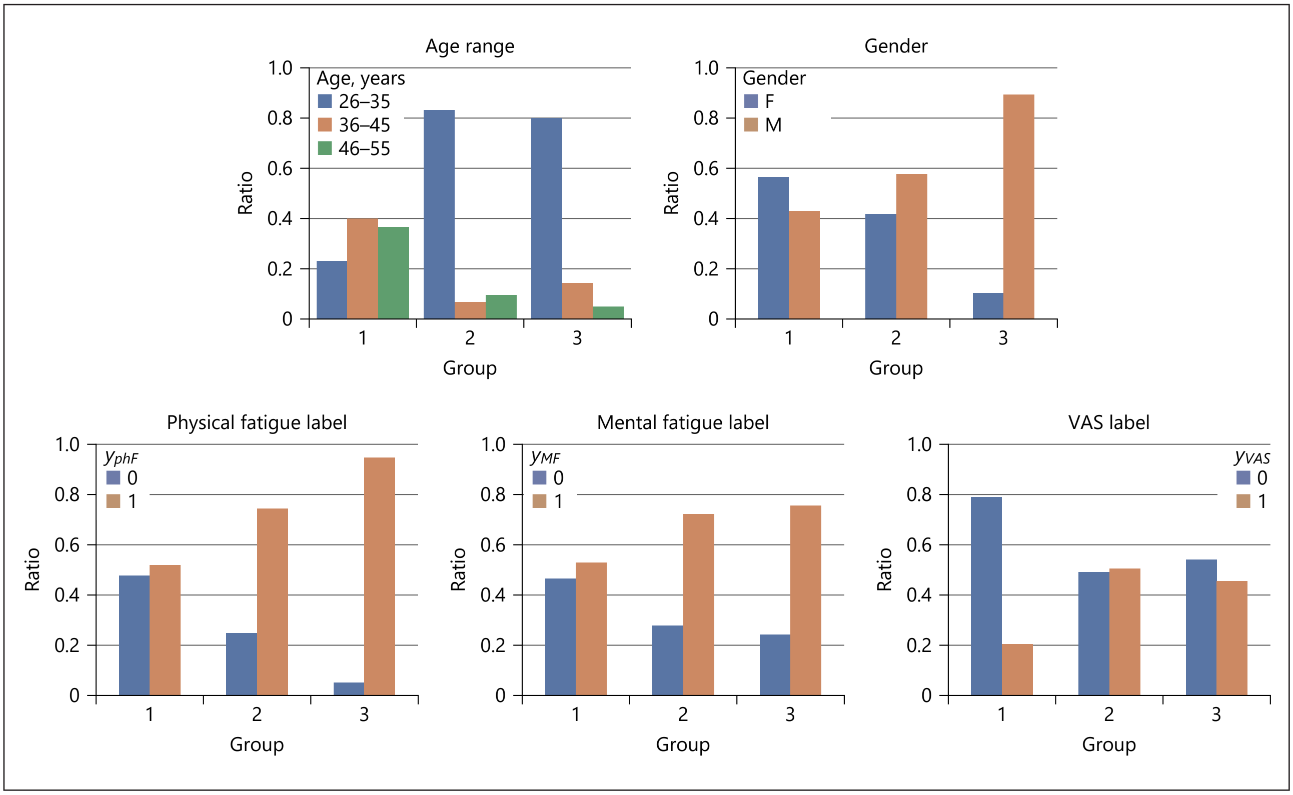

Fig. 3. Ratio of age range and gender, and physical, mental, and VAS fatigue labels among the 3 clustered groups.

tical summary features, we observed that both physical activity (ActivityCounts, EnergyExpenditure, and Steps) and vital signs (HR, HRV, and RESP) contributed to the prediction, supporting the suggestion that multimodal data are necessary to measure fatigue. In particular, the relevance of vital signs for predicting non-pathological physical fatigue supports existing findings [46–50]. Correlation of individual sensor parameters to fatigue labels demonstrated that these data, taken individually, are not informative to measure fatigue. The top 5 predictive features for physical fatigue are mainly related to physical activity (EnergyExpenditure), while for mental fatigue are mostly vital signs (RESP and HRV), suggesting that physical fatigue in healthy subjects is mainly associated with physical exhaustion, while mental fatigue is more like a psychobiological state [51].

The highest F1 score was achieved by cCNN + RF for both physical and mental fatigue prediction. We found that most of the classification error comes from observations where representations are very similar (low Euclidean distance, corresponding to similar digital data) but with different fatigue labels, potentially indicating inconsistent reporting of fatigue. We expect that increasing the diversity and size of the observation and population will improve encodings distribution and classification performance, in addition to other methods of increasing the reliability of PRO responses.

In this study, we collected data from a relatively homogeneous population of healthy adults with similar working environments. Future work should focus on collecting larger and more diversified datasets, including patients affected by pathological fatigue. Extending the monitoring time is also relevant for learning complex behavioral patterns. This would also allow for testing other state-of-the-art methodologies on multivariate time series classification, such as residual networks [52], and improve our cCNN-based representation.

Anomaly Detection

In the analyzed dataset, only 11% of the observations from 52% of subjects corresponded to experiencing fatigue (*PhF* or *MF* ∈ {*often*; *always*}), suggesting that the task of detecting fatigue could be modeled as an anomaly detection problem. Yet error rates indicate that some high fatigue observations have similar patterns and densities to low fatigue ones. Similar to the classification task, it is likely that our dataset captures a limited number of fatigue patterns due to its sample size. This promising approach takes inspiration from so-called "N-of-1" studies [53], and suggests that longer monitoring periods, including prior to symptom presentation, could improve performance.

Cluster Analysis and Digital Phenotyping

Prior to analyzing clusters from sensor data, we observed no statistically significant difference between daily averages of sensor parameters, aggregated by fatigue labels. We found that male individuals in this study had significantly higher physical activity (ActivityCounts, Steps, EnergyExpenditure) and lower mean HR, RESP, and BloodPulseWave than females, in agreement with published research [54]. We did not observe any significant difference in physical activity among age ranges. However, we found significantly higher HR, RESP, BloodPerfusion, and BloodPulseWave in the higher age range (36—55 years), consistent with previous findings on aging and vital signs [55].

We observed several digital phenotype clusters. While the interpretation of physical activity parameters (ActivityCounts, Steps, or EnergyExpenditure) is straightforward, changes in physiological data like HRV or SkinTemperature depend on other factors such as individual fitness, age, or gender. In a cluster of observations, reported physical fatigue might be a result of physical exhaustion. However, it is difficult to detect subjects' fatigue in more complex scenarios, for example in the presence of stress. Predicting fatigue intensity (VAS) was more difficult than the presence or frequency of fatigue episodes, possibly due to the subjectivity of such PRO. Baseline observations and intra-subject normalizations could improve results. Mental fatigue was less predictable than physical fatigue. Future work focusing on collecting more data sources, for example mood and cognitive assessments, will help us better understand mental fatigue [56, 57], especially in clinical trials.

Conclusion

To the best of our knowledge, this is the first work on analyzing multiple digital data sources on physical activity, vital signs, and other physiological parameters, and their link to subject-reported non-pathological physical and mental fatigue in real-world settings through multiple machine learning-based approaches: classification, clustering, and anomaly detection, thus providing different perspectives on the data. Our classification pipeline imputes missing sensor data using a recurrent neural network to compensate for the high variability of sensor wear time, learns unsupervised sensor features using a causal convolutional network, discretizes fatigue scores to binary labels to reduce intra- and intra-subject variabilities, and uses RF for predicting fatigue label from sensor features. We achieved a weighted F1 score of 0.71 ± 0.04 for predicting physical fatigue versus a lower score of 0.65 ± 0.02 for mental fatigue. Top predictive parameters came from both physical activity and vital signs, supporting the idea that fatigue is a highly multimodal concept. Cluster analysis of statistical summaries of sensor parameters resulted in a digital phenotype of physical fatigue, driven by a high intensity of physical activity. Further investigation of these digital phenotypes could help us further understand behavioral factors influencing HRQoL, for example stress. Anomaly detection showed promising results for episodic fatigue events. Our approach is based on longitudinal data collected in free-living settings and hence can be simply scaled and deployed also in clinics. The main limitation of our work is the relatively small dataset of only healthy subjects. In the future we plan on collecting more data and to retrain or transfer these solutions to clinical trial data with larger populations monitored over several months. Future directions of research will include the development of composite digital fatigue scores to quantify and augment self-reported information on this multifaceted HRQoL factor.

Acknowledgements

We thank our colleagues from Novartis: Jelena Curcic, Sarah Hamsley, Juliet Leighton-Davies, and Roland Huber for their help in setting up the study; Tom Chapuis for his suggestions and constant, constructive feedback; Carlotta Caroli, Nicholas Kelley, and Shahram Ebadollahi from the Novartis AI for Life Residency Program; Anna Kostikova, Jason Laramie, and Scott Kennedy for their support. We also thank Anika Uhde from Biovotion for her support with the Everion devices.

Statement of Ethics

Data were collected as part of a single-site, non-interventional study carried out at the Novartis Institutes for Biomedical Research (NIBR) with the project title "Assessing Technical Feasibility of Novel Digital Devices to Enable Rapid Testing and Incorporating New Technologies into Clinical Studies." Ethikkommission Nordwest- und Zentralschweiz (EKNZ), the Ethics Committee involved in the study, granted approval and written consent was given by every participant.

Conflict of Interest Statement

I.C. is an employee of and holds stock options in Evidation Health. He has received payment for lecturing on Digital Health at the ETH Zurich and FHNW Muttenz. He is an Editorial Board Member at Karger Digital Biomarkers and a founding member of the Digital Medicine Society. The other authors have no conflicts of interest to declare.

Funding Sources

This work was sponsored by Novartis and received no direct funding.

Author Contributions

H.L. and P.-A.L. contributed to data exploration and analysis, model implementation, interpretation of the results, and writing the manuscript. M.J. supervised the selection and implementation of machine learning models. I.C. and V.D.L. contributed to the conception and design of the work, data acquisition and interpretation, and drafting and revision of the manuscript. V.D.L. guided the overall project. All authors read and approved the final manuscript.

References

1 Centers for Disease Control and Prevention. Measuring healthy days: population assessment of health-related quality of life. Atlanta: CDC; 2000.

2 Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. BMJ. 2002 May;324(7347):1193–4.

3 Broderick JE, Stone AA. Paper and electronic diaries: too early for conclusions on compliance rates and their effects—Comment on Green, Rafaeli, Bolger, Shrout, and Reis (2006). Psychol Methods. 2006 Mar;11(1):106–11.

4 Dowling NM, Bolt DM, Deng S, Li C. Measurement and control of bias in patient reported outcomes using multidimensional item response theory. BMC Med Res Methodol. 2016 May;16(1):63.

5 Clay I. Impact of digital technologies on novel endpoint capture in clinical trials. Clin Pharmacol Ther. 2017 Dec;102(6):912–3.

6 Li X, Dunn J, Salins D, Zhou G, Zhou W, Schüssler-Fiorenza Rose SM, et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. PLoS Biol. 2017 Jan; 15(1):e2001402.

7 Mueller A, Hoefling H, Nuritdinow T, Holway N, Schieker M, Daumer M, et al. Continuous monitoring of patient mobility for 18 months using inertial sensors following traumatic knee injury: a case study. Digit Biomark. 2018 Aug;2(2):79–89.

8 Hannesdottir K, Sverdlov O, Curcic J, Vargas G, Hache T, Serra D, et al. P4-339: The Media Study: A Novel Method for Evaluating Digital Endpoints In Alzheimer's Disease. Alzheimer's Demen. 2006 Jul;14(7S_Part_30):P1596–7.

9 Chen R, Jankovic F, Marinsek N, Foschini L, Kourtis L, Signorini A, et al. Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM; 2019. https://doi.org/10.1145/3292500.3330690.

10 Bahej I, Clay I, Jaggi M, De Luca V. Prediction of Patient-Reported Physical Activity Scores from Wearable Accelerometer Data: A Feasibility Study. Converging Clinical and Engineering Research on Neurorehabilitation III. Springer International Publishing; 2018. pp. 668–72.

11 Galván-Tejada JI, Celaya-Padilla JM, Treviño V, Tamez-Peña JG. Multivariate Radiological-Based Models for the Prediction of Future Knee Pain: data from the OAI. Comput Math Methods Med. 2015;2015:794141.

12 Lopez-Martinez D, Picard R. Continuous pain intensity estimation from autonomic signals with recurrent neural networks. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2018. https://doi.org/10.1109/EMBC.2018.8513575.

13 Sano A, Picard RW. Stress Recognition Using Wearable Sensors and Mobile Phones. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. IEEE; 2013. DOI:https://doi.org/10.1109/ACII.2013.117.

14 Smets E, Rios Velazquez E, Schiavone G, Chakroun I, D'Hondt E, De Raedt W, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. NPJ Digit Med. 2018 Dec;1(1):67.

15 Sano A, Yu AZ, McHill AW, Phillips AJ, Taylor S, Jaques N, et al. Prediction of happy-sad mood from daily behaviors and previous sleep history. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015. https://doi.org/10.1109/EMBC.2015.7319954.

16 Finsterer J, Mahjoub SZ. Fatigue in healthy and diseased individuals. Am J Hosp Palliat Care. 2014 Aug;31(5): 562–75.

17 Kluger BM, Krupp LB, Enoka RM. Fatigue and fatigability in neurologic illnesses: proposal for a unified taxonomy. Neurology. 2013 Jan;80(4):409–16.

18 Hofman M, Ryan JL, Figueroa-Moseley CD, Jean-Pierre P, Morrow GR. Cancer-related fatigue: the scale of the problem. Oncologist. 2007;12(S1 Suppl 1):4–10.

19 Chaudhuri A, Behan PO. Fatigue in neurological disorders. Lancet. 2004 Mar;363(9413):978–88.

20 Mollayeva T, Kendzerska T, Mollayeva S, Shapiro CM, Colantonio A, Cassidy JD. A systematic review of fatigue in patients with traumatic brain injury: the course, predictors and consequences. Neurosci Biobehav Rev. 2014 Nov;47:684–716.

21 Zielinski MR, Systrom DM, Rose NR. Fatigue, Sleep, and Autoimmune and Related Disorders. Front Immunol. 2019 Aug;10:1827.

22 Aaronson LS, Pallikkathayil L, Crighton F. A qualitative investigation of fatigue among healthy working adults. West J Nurs Res. 2003 Jun;25(4):419–33.

23 Waterhouse J, Edwards B, Nevill A, Atkinson G, Reilly T, Davies P, et al. Do subjective symptoms predict our perception of jet-lag? Ergonomics. 2000 Oct;43(10):1514–27.

24 Waterhouse J, Reilly T, Atkinson G, Edwards B. Jet lag: trends and coping strategies. Lancet. 2007 Mar; 369(9567):1117–29.

25 Aryal A, Ghahramani A, Becerik-Gerber B. Monitoring fatigue in construction workers using physiological measurements. Autom Construct. 2017;82:154–65.

26 Sedighi Maman Z, Alamdar Yazdi MA, Cavuoto LA, Megahed FM. A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. Appl Ergon. 2017 Nov;65:515–29.

27 Maman ZS, Chen YJ, Baghdadi A, Lombardo S, Cavuoto LA, Megahed FM. A data analytic framework for physical fatigue management using wearable sensors. Expert Syst Appl. 2020;155:113405.

28 De Beéck TO, Meert W, Schütte K, Vanwanseele B, Davis J. "Fatigue prediction in outdoor runners via machine learning and sensor fusion." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018. https://doi.org/10.1145/3219819.3219864.

29 Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Min Knowl Discov. 2019 Mar;33(4):917–63.

30 Franceschi JY, Dieuleveut A, Jaggi M. Unsupervised scalable representation learning for multivariate time series. In Advances in Neural Information Processing Systems. 2019. pp. 4650–61.

31 Jaques N, Taylor S, Sano A, Picard R. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE; 2017. https://doi.org/10.1109/ACII.2017.8273601.

32 Michielsen HJ, De Vries J, Van Heck GL. Psychometric qualities of a brief self-rated fatigue measure: The Fatigue Assessment Scale. J Psychosom Res. 2003 Apr;54(4):345–52.

33 Lee KA, Hicks G, Nino-Murcia G. Validity and reliability of a scale to assess fatigue. Psychiatry Res. 1991 Mar; 36(3):291–8.

34 Cao W, Wang D, Li J, Zhou H, Li L, Li Y. Brits: Bidirectional recurrent imputation for time series. In Advances in Neural Information Processing Systems. 2018. pp. 6775–85.

35 Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. 2013. pp. 3111–9.

36 Kotronoulas G, Kearney N, Maguire R, Harrow A, Di Domenico D, Croy S, et al. What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. J Clin Oncol. 2014 May; 32(14):1480–501.

37 Khanna D, Pope JE, Khanna PP, Maloney M, Samedi N, Norrie D, et al. The minimally important difference for the fatigue visual analog scale in patients with rheumatoid arthritis followed in an academic clinical practice. J Rheumatol. 2008 Dec;35(12):2339–43.

38 Kos D, Nagels G, D'Hooghe MB, Duportail M, Kerckhofs E. A rapid screening tool for fatigue impact in multiple sclerosis. BMC Neurol. 2006 Aug;6(1):27.

39 Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data–recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction. 2013. pp. 245–51.

40 Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000. pp. 93–104.

41 Ward JH Jr. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963 Mar;58(301):236–44.

42 De Luca V, Luo H, Clay I. Continuous multi-sensor wearable data and daily subject-reported fatigue of heathy adults [Data set]. Zenodo; 2020. http://doi.org/10.5281/zenodo.4266157

43 Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012 Sep;24(3):69–71.

44 Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 1998 Sep;10(7):1895–923.

45 Nelesen R, Dar Y, Thomas K, Dimsdale JE. The relationship between fatigue and cardiac functioning. Arch Intern Med. 2008 May;168(9):943–9.

46 Samel A, Wegmann HM, Vejvoda M. Aircrew fatigue in long-haul operations. Accid Anal Prev. 1997 Jul;29(4):439–52.

47 Tran Y, Wijesuriya N, Tarvainen M, Karjalainen P, Craig A. The relationship between spectral changes in heart rate variability and fatigue. J Psychophysiol. 2009 Jan;23(3):143–51.

48 Schmitt L, Regnard J, Millet GP. Monitoring fatigue status with HRV measures in elite athletes: an avenue beyond RMSSD? Front Physiol. 2015 Nov;6:343.

49 Escorihuela RM, Capdevila L, Castro JR, Zaragozà MC, Maurel S, Alegre J, et al. Reduced heart rate variability predicts fatigue severity in individuals with chronic fatigue syndrome/myalgic encephalomyelitis. J Transl Med. 2020 Jan;18(1):4.

50 Muñoz JE, Gouveia ER, Cameirão MS, Badia SB. i Badia SB. PhysioLab-a multivariate physiological computing toolbox for ECG, EMG and EDA signals: a case of study of cardiorespiratory fitness assessment in the elderly population. Multimedia Tools Appl. 2018 May;77(9):11521–46.

51 Mizuno K, Tanaka M, Yamaguti K, Kajimoto O, Kuratsune H, Watanabe Y. Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. Behav Brain Funct. 2011 May;7(1):17.

52 Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: A strong baseline. 2017 International Joint Conference on Neural Networks (IJCNN). IEEE; 2017. https://doi.org/10.1109/IJCNN.2017.7966039.

53 Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med. 2011 Mar;8(2):161–73.

54 Lutfi MF, Sukkar MY. The effect of gender on heart rate variability in asthmatic and normal healthy adults. Int J Health Sci (Qassim). 2011 Jul;5(2):146–54.

55 Chester JG, Rudolph JL. Vital signs in older patients: age-related changes. J Am Med Dir Assoc. 2011 Jun;12(5):337–43.

56 Marcora SM, Staiano W, Manning V. Mental fatigue impairs physical performance in humans. J Appl Physiol (1985). 2009 Mar;106(3):857–64.

57 Valentine AD, Meyers CA. Cognitive and mood disturbance as causes and symptoms of fatigue in cancer patients. Cancer. 2001 Sep;92(6 Suppl):1694–8.