PNAS Nexus, 2024, 3, pgae421
https://doi.org/10.1093/pnasnexus/pgae421
Advance access publication 15 October 2024
Research Report

Wearable network for multilevel physical fatigue prediction in manufacturing workers

Payal Mohapatra[a,†], Vasudev Aravind[b,†], Marisa Bisram[b], Young-Joong Lee[g], Hyoyoung Jeong[g], Katherine Jinkins[g], Richard Gardner[c], Jill Streamer[d], Brent Bowers[e], Lora Cavuoto[f], Anthony Banks[h], Shuai Xu[h,i], John Rogers[a,b,g,h], Jian Cao[b], Qi Zhu[a,*] and Ping Guo[b,*]

[a]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA
[b]Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA
[c]Boeing Research & Technology, Everett, WA 98203, USA
[d]Boeing Research & Technology, Ladson, SC 29456, USA
[e]Global Occupational Safety, Deere and Company, Moline, IL 61265, USA
[f]Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY 14260, USA
[g]Department of Biomedical Engineering, Northwestern University, Evanston, IL 60208, USA
[h]Querrey Simpson Institute for Bioelectronics, Northwestern University, Evanston, IL 60208, USA
[i]Sibel Health Inc., Chicago, IL 60614, USA
*To whom correspondence should be addressed: Email: qzhu@northwestern.edu (Q.Z.); ping.guo@northwestern.edu (P.G.)
†P.M. and V.A. contributed equally to this work.
Edited By: Xiaowei Yue

Abstract

Manufacturing workers face prolonged strenuous physical activities, impacting both financial aspects and their health due to work-related fatigue. Continuously monitoring physical fatigue and providing meaningful feedback is crucial to mitigating human and monetary losses in manufacturing workplaces. This study introduces a novel application of multimodal wearable sensors and machine learning techniques to quantify physical fatigue and tackle the challenges of real-time monitoring on the factory floor. Unlike past studies that view fatigue as a dichotomous variable, our central formulation revolves around the ability to predict multilevel fatigue, providing a more nuanced understanding of the subject's physical state. Our multimodal sensing framework is designed for continuous monitoring of vital signs, including heart rate, heart rate variability, skin temperature, and more, as well as locomotive signs by employing inertial motion units strategically placed at six locations on the upper body. This comprehensive sensor placement allows us to capture detailed data from both the torso and arms, surpassing the capabilities of single-point data collection methods. We developed an innovative asymmetric loss function for our machine learning model, which enhances prediction accuracy for numerical fatigue levels and supports real-time inference. We collected data on 43 subjects following an authentic manufacturing protocol and logged their self-reported fatigue. Based on the analysis, we provide insights into our multilevel fatigue monitoring system and discuss results from an in-the-wild evaluation of actual operators on the factory floor. This study demonstrates our system's practical applicability and contributes a valuable open-access database for future research.

Keywords: wearable sensors, quantifying physical fatigue, continuous fatigue monitoring, real-time machine learning, manufacturing

Significance Statement

This research enhances occupational health by improving ergonomics and managing fatigue among manufacturing workers. Leveraging advanced multimodal wearable sensors and lightweight machine learning algorithms, it enables continuous, real-time fatigue monitoring. The study addresses limitations in adaptive sensing technologies and explores complex biomarker-fatigue relationships. Data from 43 participants in diverse manufacturing tasks reveal insights such as the impact of nondominant arm kinetics on fatigue, and the role of body mass, age, and gender. The research also highlights the significance of physiological signs in fatigue perception and confirms that fatigue characteristics are highly personalizable, with better prediction performance for users whose data were included in training.

Introduction

Manufacturing workplaces are physically strenuous, with US employers incurring an estimated annual cost of $136 billion due to health-related loss in productivity in the manufacturing sector (1). Surveys consistently show a high prevalence of fatigue among manufacturing workers in Canada (2), the EU (3), Japan (4), and

Competing Interest: The authors declare no competing interests.
Received: April 11, 2024. **Accepted:** September 9, 2024

Sweden (5). A meta-analysis of fatigue research indicates that 90% of shift workers regularly experience fatigue and sleepiness during work (6). Fatigue not only increases the risk of injuries and accidents, causing loss of production, but also leads to significant health issues, including discomfort, reduced motor control (7, 8), musculoskeletal disorders (9), chronic fatigue syndrome (10), and diminished immune functions (4). The economic loss of chronic fatigue syndrome alone is estimated at $17 to $24 billion annually (11).

There are no universally recognized biomarkers to quantify physical fatigue, making the task of fatigue monitoring in manufacturing workers extremely hard. Some studies (12) view physical fatigue as a rule-based target—Rapid Upper Limb Assessment (RULA) and Rapid Entire Body Assessment (REBA) scores (13–15)—based on the pose estimates of the operator. Moreover, fatigue is not an isolated musculoskeletal outcome but a byproduct of exhaustion as well. Perceived fatigue scores using Borg ratings (16) are more accurate. Many studies utilize the standardized Perceived Stress Scale (PSS) (17, 18) and Trier Social Stress Test (TSST) (19) to measure psychological stress with absolute labels for fatigue and nonfatigue (based on the stress test protocol). Other domains use TSST-like settings to arrive at a Borg rating of perceived exertion (RPE). Although Maman et al. (20) used multiscale fatigue, while for reporting, they ultimately converted it to a binary state for final prediction. Another study (21), used RPE to quantify physical workload by tracking energy expenditure and compared the physiological and perceptual responses between bar benders and bar fixers when they worked in a hot environment. Most, if not all, of the previous works view fatigue as a dichotomous variable which may be an oversimplification of physical exertion. There is limited consensus in either the sports science (22) literature or manufacturing applications (23) that fatigue should be viewed as a continuous variable. Some works (24) also assume that the worker starts from a nonfatigued state and the prediction model is used to determine a deviation from this baseline.

A focal technique for endocrine stress measurement when subjected to psychosocial stressors involves salivary cortisol (25, 26) measurement. Extending this momentary measurement to indicate physical fatigue in operators on factory floors is impractical. Most pose-estimated physical fatigue indicators, such as RULA and REBA scores (13–15), rely on visual sensing for accurate skeletal tracking, which can intrude on worker privacy. Using self-reported fatigue scores with wearable sensing has emerged as a more favored framework for monitoring fatigue unobtrusively. Physical fatigue labels are more challenging on account of the subjectivity of fatigue per individual, especially under realistic uncontrolled manufacturing protocols. Some fatigue-sensing approaches rely on wrist-based wearables (27), which only provide a single point of measurement and can be obtrusive in manufacturing settings since most operators wear gloves for their tasks.

In the manufacturing and ergonomics literature, many works (24, 28) base their *fatigue-sensing* on gait characterization or lower-limb kinematics under the assumption of significant walking during an operator's shift. Zhang et al. (29) achieved an accuracy of 94% in a fatigue monitoring study using inertial motion units (IMUs) and a support vector machine (SVM) based learner for a bricklaying task. The hypothesis was that jerk values can be compared between rested and fatigued states to provide a fatigue scale. Their study used only expert masons for the task and hence nothing can be said for the applicability of the system to a wider demographic. We move away from this assumption and focus on two very different manufacturing tasks and study task-specific metrics that can indicate operator fatigue.

At a high level, physical fatigue lacks universally accepted metrics and distinct biomarkers. Considering previous works in fatigue monitoring and the practical challenges of implementing a continuous fatigue monitoring system in manufacturing settings, we identify several prominent research gaps: (i) limited expressiveness of physical fatigue states, resulting in oversimplified characterizations, (ii) absence of multimodal, practical sensing and analytics frameworks that are suitable for deployment on factory floors, and (iii) a shortage of publicly available datasets for studying physical labor-induced fatigue.

To address the first challenge of the constrained depiction of fatigue states, we view an operator's fatigue state as a continuous hidden variable. We hypothesize that with multimodal sensing capabilities, we can capture physiological and kinematic compensation when an operator develops *fatigue*. To provide meaningful feedback to the operators and manufacturing companies, we support a more granular fatigue scale compared to other studies (30, 31). Consequently, we measure fatigue based on perceived exertion (RPE), which is commonly used in many sports science studies. We specifically use a linear projection of the Borg Scale between 0 and 10. In addition to introducing fatigue as a continuous target variable, we integrate a custom asymmetric objective into our machine-learning approach, emphasizing an operator-centered approach to fatigue modeling.

Secondly, we explore multimodal wearable sensors that are flexible and skin-compatible for continuous wear. With efficient learning techniques and the different *views* for fatigue markers, we aim to enable timely prevention to relieve fatigue by making predictions close enough to human perception. Our prediction model makes use of data from an unobtrusive sensing system that utilizes state-of-the-art soft wearable wireless sensors (ANNE and ADAM sensors) first introduced by Chung et al. in (32, 33). The wearable system is time-synchronized and can be controlled using a central hub streaming real-time vital signs including electrocardiogram (ECG), heart rate (HR), skin temperature, and locomotive signals from the accelerometers and gyroscopes.

Finally, we gather data under realistic conditions by replicating two strenuous manufacturing tasks across 43 participants. We perform comprehensive characterization and validation through large-scale data collection in a laboratory setting, which will be made publicly available to promote further research in this area. Additionally, we demonstrate our method on two factory floors. We further interpret our machine learning model and provide insights into the features that influence its prediction under various settings. We also report the user feedback on our methodology through ecological momentary assessment (EMA) (17). Our approach combines multimodal sensing with a data analytics framework and near-real-time visualization to predict the multilevel fatigue of manufacturing operators.

Results

In this section, we briefly present the design of our study followed by a discussion on the feature analysis of predicting fatigue using different sensing modalities. We then present our regression-based formulation to predict continuous perceived fatigue using an asymmetric loss function with an average mean absolute error of **2.27** across multiple tasks and users on a continuous fatigue scale of 0–10. The overall study framework is illustrated in Fig. 1. Note that our target for prediction is a perceived rating of exertion (modified Borg scale (34)), which is a subjective variable. Studies (35, 36) have shown that applications (like emotion state,

sports science, rehabilitation, etc.) relying on self-reported subjective scores tend to be inherently noisy. Objectively quantifying the bias, risk, and subjectivity of a score is highly application and demographic-specific. With this noise in consideration, we certify the validity of our method by deploying them in-the-wild and across tasks. We also conduct ablation studies to show the impact on the nondominant arm as a prominent indicator of fatigue state across tasks along with individual-specific attributes like age, gender, height, weight, etc.

Study overview

The goal of the experiment is to predict fatigue trends in a subject, while they are asked to perform predefined manual tasks simulating a manufacturing environment, using data from soft, flexible, wearable sensors, and a vision system. The tasks in this study are repetitive and physically exerting, involving intricate steps taken in real manufacturing settings. The iterative nature of the tasks facilitates comparative analyses of distinct temporal segments to characterize fatigue. The two manufacturing tasks are (i) Task Composite: Composite Sheet Layup and (ii) Task Harnessing: Wire Harnessing. A Movie S1 is provided to highlight the experimental protocols and task designs. Details regarding the test bed dimensions can be found in Figs. S1 and S2. The task protocol requires the subject to wear sensors to monitor vital and locomotive signs continuously. Additionally, we incorporate a weighted vest to exaggerate the induced fatigue in a reasonable duration for the study to mimic a full shift for a manufacturing worker. Each task consists of two rest periods of 5 min each at the start and end, as well as five segments of physical tasks. On average, each task takes a total of one hour. Before each data segment, the subject fills out a survey form to indicate their current fatigues as per the Borg scale.

Summary of predictive models

We formulate fatigue as a continuous variable and construct a regression model that can penalize incorrect predictions based on their exact error from a reported score. To handle a complex task in a data-poor setting we employ gradient-boosted trees-based regressor and a custom asymmetric objective for optimization. Our feature set is constructed by fusing the vital and locomotive sensor data along with person-specific attributes (like age, gender, height, weight, etc.). We analyze two tasks and study the transferability across tasks, individuals, and the effect of various sensing modalities.

Data statistics from user study

We carried out extensive data collection at Northwestern University for 18 months under Institutional Review Board (IRB) approval (STU0021461) and endorsement by the US Army Human Research Protections Office (MXD191305). We collected data from 43 participants. After quality assessment (as shown in Analysis framework section), we obtained usable data from 41 participants for the Composite Task and 36 participants for the Harnessing Task. The participants ranged from the age group 18–56 years old. Both the tasks have nine female participants representing about 23.7% of the total data. Figure S6 illustrates the self-reported fatigue scores by the participants during each segment of the tasks on a scale of 0–10. In Table 1, the ranges and average values of user statistics (age, weight, height, and gender) are reported. The average age of the participants from Northwestern University is 24.83, with four individuals aged 29 or above. The oldest participant was aged 56 years. The weight range for participants lies between 96–220 lbs, with an average value of 154.09 lbs. The average height of the participants was 173.05 cm. This study had a gender ratio (m/f) of 3.9. Most of the participants were university students with a few participants from the university workshop.

Table 1. Summary of user-defined statistics including ranges and average values for age, weight, height, and gender ratio (m/f) of all Northwestern University participants.

User statistics	Average	Max	Min
Age (years)	24.83	56	18
Weight (lbs)	154.09	220	96
Height (cm)	173.05	190	151
Gender ratio (m/f)	3.90		

We derive various features from the preprocessed sensor data. To study their effects on perceived physical fatigue, we first use principal component analysis (PCA) (37). Examining the magnitude of the eigenvectors indicates the feature-importance order as shown in Fig. 2 for a combination of Task Composite and Task Harnessing. The features are categorized as (i) person-specific factors (gender, age, weight, and height), (ii) physiological signs (heart rate, heart rate variability, and skin temperature), and (iii) IMU-related signs (accelerometer, gyroscope on five locations). The PCA analysis indicates that features based on an operator's personal attributes and physiological signs are relatively more important across both tasks. Figures S8 and S9 illustrate explicit feature importance for the composite and harnessing tasks respectively.

Next, we compute Spearman's correlation matrix which assesses the strength and direction of monotonic association between two variables (38). Figures S10–S12 depict the correlation analyses of the standardized feature space for the Task Composite, Task Harnessing, and a combination of Composite and Harnessing tasks, respectively. The highest value for $|\rho|$, Spearman's correlation coefficient across both tasks for a random variable $X \in$ *Feature Space* and $Y =$ *Fatigue Label*, is 0.25 in Task Composite (Fig. S10) for average skin temperature and 0.29 in Task Harnessing (Fig. S11) for cardiac capacity. Across all the task settings, we consistently observe that the velocity of the ADAM5 sensor on the chest (rms_imu5_vel) has a negative relation with the fatigue variable, which could indicate an overall slowness in the operator as fatigue levels rise. We observe $\rho >$ 0.7 for features *_imu1_* and *_imu3_* in Task Composite in Fig. S10 corresponding to right upper and left upper arm ADAM sensors, respectively. This can be attributed to the rhythmic/symmetric nature of this task. The Spearman correlation between average heart rate (avg_hr) and average heart rate variability (avg_hrv) for the Task Composite, Harnessing, and the combined dataset are 0.15, −0.45, and 0.001, respectively. Past works have supported such observations that under parasympathetic or sympathetic stressful conditions there is an increase in heart rate and a decrease in heart rate variability (39, 40).

Data statistics from performance tracking

Fatigue can negatively affect performance. Thus, tracking performance can give valuable insights into fatigue effects and how these effects can be curbed. To this end, the tasks designed also have a scoring metric defined to allow performance tracking. This can be used to check for periods where the quality of the job is suffering and further understand the role of fatigue in the decrease in operator performance.

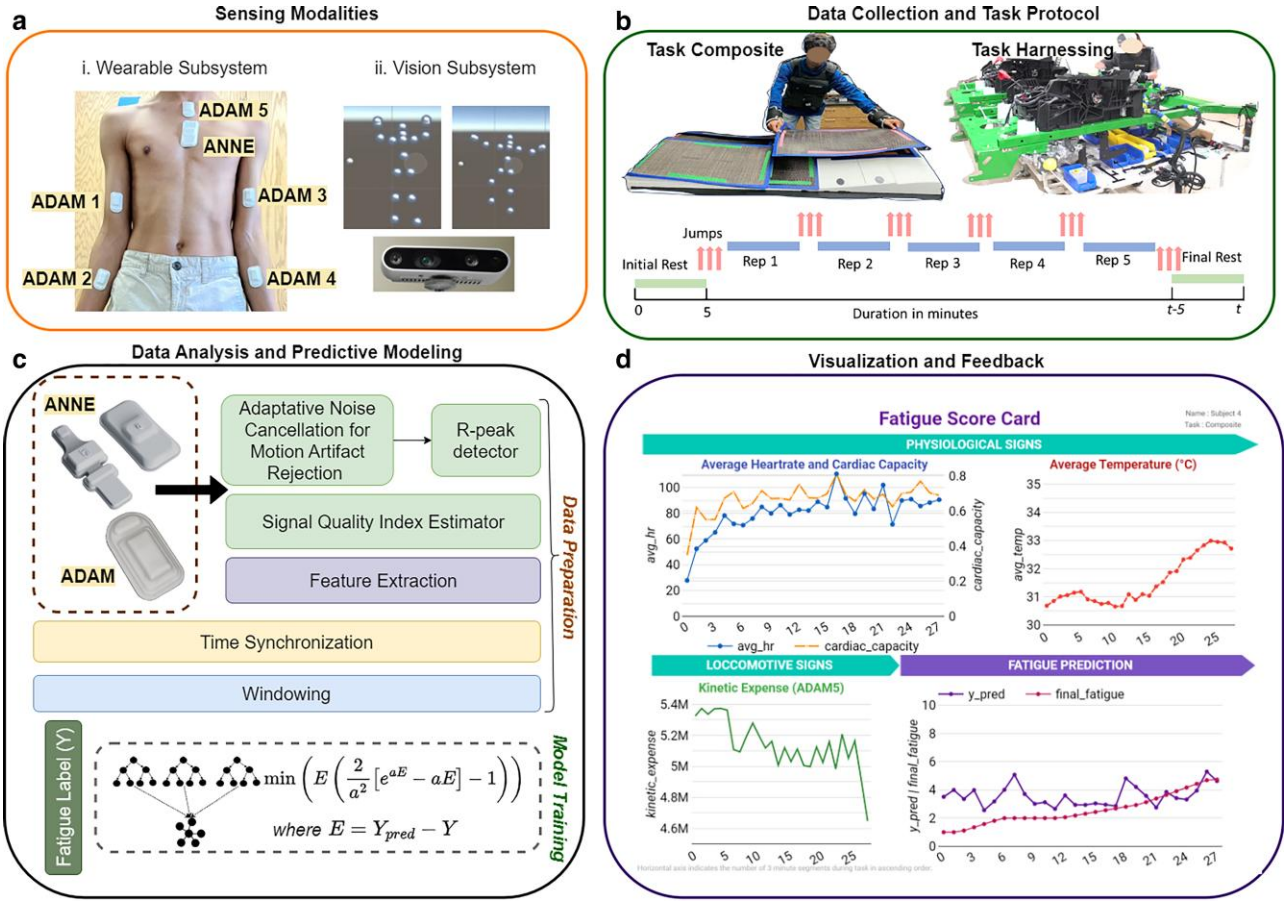

Fig. 1. Overview of the system design. a) (i) Wearable sensor data snapshots and sensor locations are indicated on the operator. (ii) The auxiliary vision system tracking the skeleton of operator motions. b) Composite sheet layup (Task Composite) and wire harnessing (Task Harnessing) setups are shown with task protocols. c) Overall machine learning pipeline—data preprocessing, ECG denoising, data filtering, and windowing, followed by model training and evaluation. d) A mock-up of the visualization dashboard, which gives near-real-time feedback to the operator.

The performance metrics are defined for the two designed manufacturing tasks as follows. For Task Composite, performance can be judged by sheet placement, orientation, and quality of smoothing. Line detection is used to pick out the sheet's position relative to the mold. The scoring script then checks the placement of each sheet and provides a score for the same. Details regarding performance metrics for Task Composite can be found in Fig. S3. For Task Harnessing, the quality of the job would be determined by the tightness and positioning of the zip ties within the predefined zones along the cable. We measure the length of the tail end of the zip tie after it has been tightened. Participants were asked to trim the zip tie after tightening. The length of the trimming is measured and compared to a benchmark value. The trim length relates to the effort put into applying that zip tie.

Figure 3 presents the performance scores obtained by participants for the two tasks. The scores tend to improve around Rep 2 or 3 for both tasks which can be explained partly due to "task learning." Subjects adapt to the requirements of the unseen repetitive task after a couple of tries. However, the average scores tend to drop around Rep 4 and Rep 5 which may be due to the fatiguing nature of the task after a certain period. The standard deviation of repetition scores across all subjects increases from Rep 1 to Rep 5. This is an indicator of the difference in fatiguing tendencies across different subjects. All subjects start from a well-rested state (low variance in early repetition scores) and move towards a fatigued state. Some subjects may tire more easily than others resulting in higher score variations towards the later repetitions. Another observation made is the range of scores obtained for both tasks varies drastically. Scores remain nearly consistent for Task Composite, whereas a wide range of performance scores is obtained for Task Harnessing.

Evaluation of predictive model across tasks

The distribution of physiological signs varies across individuals while also having a temporal shift for the same individual. Since our task duration is generally one hour we assume no major behavioral changes due to factors other than fatigue. We can extend this assumption to distinguish different individuals being markers of distinct physiological distributions. The distribution of the locomotive signs is dependent on both the individual as well as the nature of the task. We evaluate our model performance under two major categories—performance on unseen individuals and performance on unseen tasks.

One of the pivotal design choices in our study is the adoption of a custom asymmetric loss function, which strategically penalizes under-predictions more severely than over-predictions. This approach is specifically tailored to align with our design philosophy of prioritizing operator-centric considerations in manufacturing applications, where minimizing underprediction errors holds critical importance. We illustrate an instance of our prediction in Fig. 4 where the number of underpredictions by the same model

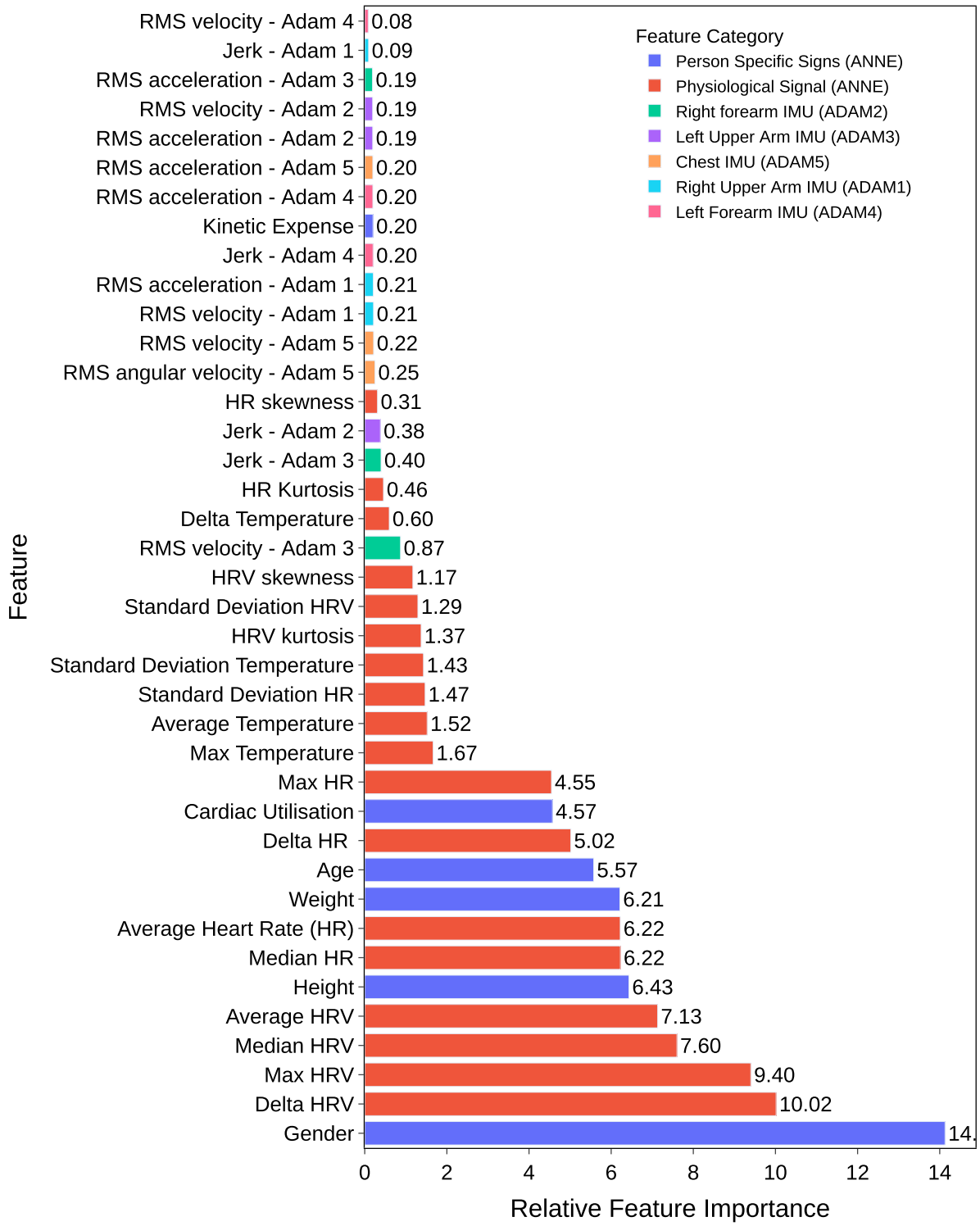

Fig. 2. Feature importance using PCA for the combination of Task Composite and Task Harnessing. The acronyms are explained as RMS, root mean square; HR, heart rate; HRV, heart rate variability. The details of ADAM and ANNE sensors are given in the Sensing framework section. For brevity, we refer ADAM sensor locations as LL: left lower arm; RL, right lower arm; RU, right upper arm; LU, left upper arm.

is reduced due to our adopted asymmetric objective. In Table 2, we report the complete performance statistics by running each experiment with three different seeds for three metrics: root mean squared error (RMSE), mean absolute error (MAE), and the underprediction rate (UR).

Our custom asymmetric LINEX objective achieves an average underprediction rate of 0.38, significantly outperforming the symmetric objective's rate of 0.71. Notably, our approach not only meets the operational target of maintaining an underprediction rate below 0.5 but also exhibits a 28% improvement in the MAE scores on average over the symmetric counterpart. As shown in Fig. S6, the distribution of self-reported labels is not uniform. Our hypothesis posits that the imposition of intensified penalization for over-predictions compels the model to acquire knowledge about patterns within regions exhibiting elevated levels of fatigue, despite the availability of comparatively limited data for such

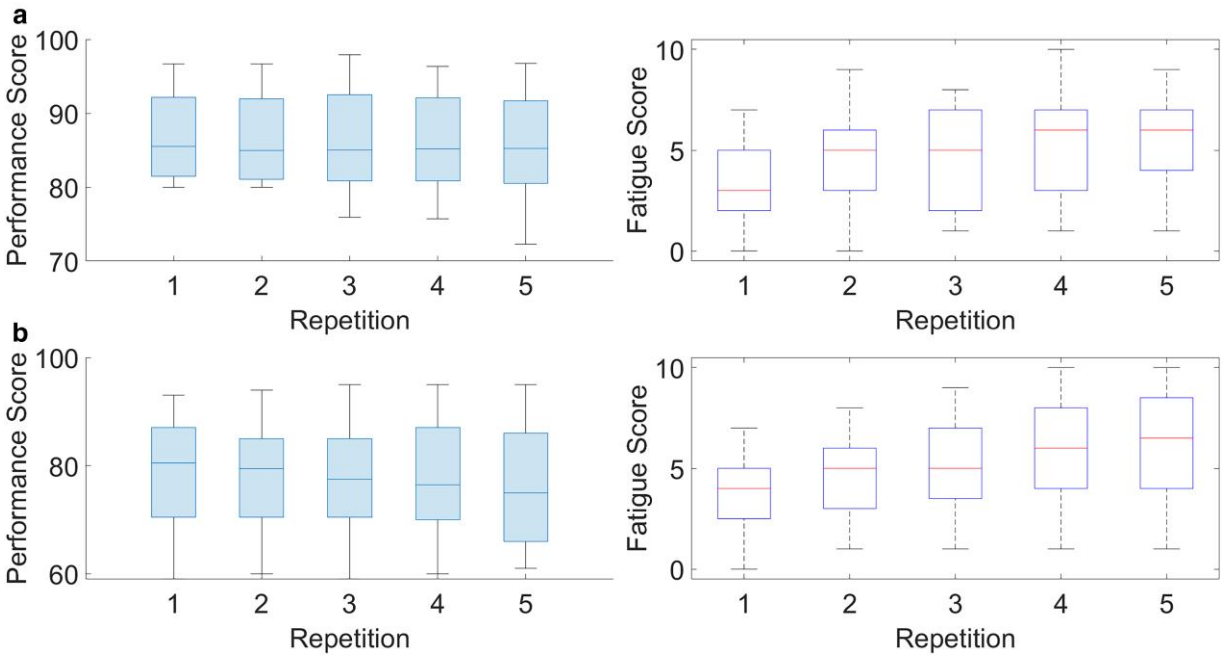

Fig. 3. Performance and fatigue score distribution across each repetition for both a) Task Composite and b) Task Harnessing.

regions. We also observe that across all task settings, there is an average of 21% boost in performance when the users being trained and tested on are overlapping. This aligns with the hypothesis that the individual operator's distribution can be very different and transferability of an existing trained model to a new operator may suffer. This observation also assures that if some training samples are provided for an operator it can perform well on unseen samples of different fatigue stages of the same operator (in this case with an average mean absolute error of 2.44). We also observed that training a model on Task Harnessing and transferring it zero-shot (without any fine-tuning) to nonoverlapping users performing Task Composite samples yields consistent performance comparable to in-domain data, while, transferring from Task Composite to Task Harnessing shows a slight decrease in performance (by 3.25%). Generally, across the overlapping and nonoverlapping users' test set, we observe the minimum average MAE of 2.29 while optimizing an asymmetric objective in the case of combined tasks using only physiological sensing features (ANNE sensor). This could be due to two reasons: (i) a larger training dataset due to combined tasks and (ii) multiple tasks allow the model to learn generalizable fatigue patterns which help in predicting fatigue of unseen targets.

Model explainability and interpretability

Analysis of the importance of sensing modalities

We want to study the impact of every sensing modality on the predictive abilities of the model. Consistent with our previous definitions of modalities, we consider physiological signs, person-specific features, and the five locations of the locomotive features. We evaluate the model performance by systematically leaving each modality out and training the model. Figure 5 shows this evaluation across the two different tasks and their combination under nonoverlapping user settings. We leave the same users across all settings for consistency.

We observed a consistent decrease in performance across all three task settings when the left upper and lower arm features were removed, resulting in an average drop of 4.76%. This contrasts with an 8% decrease specifically in the harnessing task when both the upper and lower right arm features were removed. This disparity suggests that the movements of the nondominant arm may universally signal fatigue, independent of the specific task being performed.

Furthermore, we observe that physiological indicators such as heart rate, skin temperature, and heart rate variability exert a greater influence on Task Harnessing compared to Task Composite. This distinction may stem from the unstructured nature of movements in Task Harnessing, which diminishes the predictive power of limb-related features in indicating fatigue. It is noteworthy that removing a sensing modality often results in decreased model performance for Task Harnessing, indicating the necessity of multiple modalities to effectively capture fatigue indicators in more complex tasks. Additionally, person-specific characteristics significantly impact model performance.

The ADAM sensor, located on the chest, measures the operator's overall locomotion. In this analysis, the model shows a 2.2% performance improvement in the combination of Task Composite and Harnessing. Specifically, Task Harnessing exhibits a 6.3% improvement compared to Task Composite's 1.4%, suggesting a potentially more critical role of the chest-located IMU for certain tasks over others.

Shapley scores

Understanding the inner workings of machine learning models is crucial, particularly when applying to human-centric decision-making. Due to the limited dataset and small machine learning models in this application, we can objectively explain much of the model behavior using the state-of-the-art model explanation tool called SHapley Additive exPlanations (SHAP) (41).

This tool offers explanations that are independent of the specific model employed and utilizes all samples to generate model explanations. We computed the SHAP values for each feature of every subject and visualized them as dots in an information-dense summary plot as shown in Fig. 6. In this plot, each dot represents both the true feature value and the corresponding SHAP value.

Table 2. Summary of performance of the predictions from the model vs. self-reported scores under different experimental settings.

Test distributions	Overlapping users						Nonoverlapping users					
Optimization objective	Symmetric			Asymmetric			Symmetric			Asymmetric		
Tasks	RMSE	MAE	UR	RMSE	MAE	UR	RMSE	MAE	UR	RMSE	MAE	UR
C → C	3.94 ± 0.92	3.17 ± 0.51	0.65 ± 0.11	2.52 ± 0.20	2.07 ± 0.05	0.40 ± 0.01	4.06 ± 0.11	3.10 ± 0.14	0.59 ± 0.14	2.90 ± 0.03	2.45 ± 0.04	0.44 ± 0.04
H → H	3.79 ± 1.22	2.91 ± 0.70	0.71 ± 0.16	2.68 ± 0.26	2.28 ± 0.28	0.33 ± 0.08	4.02 ± 0.68	3.47 ± 0.50	0.72 ± 0.08	2.84 ± 0.17	2.29 ± 0.12	0.38 ± 0.06
C + H → C + H	2.76 ± 0.35	2.25 ± 0.28	0.61 ± 0.11	2.77 ± 0.18	2.30 ± 0.09	0.32 ± 0.01	5.90 ± 0.91	5.19 ± 0.51	0.77 ± 0.17	2.73 ± 0.05	2.41 ± 0.03	0.40 ± 0.08
C + H → C + H (ANNE only)	4.11 ± 0.69	3.45 ± 0.84	0.88 ± 0.11	2.57 ± 0.05	2.07 ± 0.05	0.37 ± 0.06	3.13 ± 0.31	2.76 ± 0.27	0.61 ± 0.07	2.69 ± 0.41	2.50 ± 0.08	0.39 ± 0.04
C → H	2.76 ± 0.35	2.25 ± 0.28	0.61 ± 0.11	2.77 ± 0.18	2.30 ± 0.09	0.32 ± 0.01	3.83 ± 0.56	3.41 ± 0.59	0.85 ± 0.10	2.81 ± 0.11	2.54 ± 0.10	0.38 ± 0.06
H → C	4.11 ± 0.69	3.45 ± 0.84	0.88 ± 0.11	2.57 ± 0.05	2.07 ± 0.05	0.37 ± 0.06	3.81 ± 0.68	2.93 ± 0.46	0.54 ± 0.22	2.76 ± 0.12	2.44 ± 0.08	0.41 ± 0.05

RMSE, root-mean-square error; MAE, mean absolute error; UR, underprediction rate.

The true feature value is indicated by a color map, with blue representing lower values and red representing higher values. The x-axis position of a dot represents the SHAP value. The absolute SHAP value reflects the relative importance of the feature, and a positive SHAP value (on the right side of the x-axis) suggests a tendency for a positive prediction by the model, while a negative value suggests the opposite. Figure 6 ranks all the features based on the average of the absolute SHAP values from all the dots, indicating their importance. Our model demonstrates a higher fatigue prediction with a higher value of maximum heart rate in a given window. Also, the chest ADAM sensor (ADAM5) RMS angular velocity for a window is ranked as one of the top features in terms of feature importance, which aligns with the hypothesis that this modality captures the overall ambulatory intensity of the subject. Previous literature (42) shows that jerk is an important metric in inducing fatigue in manufacturing tasks specifically. Our model agrees with this trend. Moreover, our model shows that two of the features from the left lower arm contribute to the top features. This is consistent with our previous observation that features from nondominant hands indeed reflect fatigue-inducing patterns. Additional explainer plots specific to individual tasks are shown in Fig. S14.

In-the-wild User Feedback Study

The primary target users of this technology are the operators on the factory floor. The factory's decision makers can make use of this information to take necessary operational steps in the form of interventions or adaptive work schedules. To study the relevance of our methodology, we conducted two studies in large manufacturing factories in the Midwest and West Coast for each of the tasks. We tested the model trained on data from subjects at Northwestern University for their transferability on actual factory workers whose age group and skillset vary from the trained population. The task setups used are actual manufacturing units instead of the mock-up version used at the Northwestern facility. The task protocols were slightly modified (fewer repetitions, camera positioning, etc.) to adhere to factory norms and accommodate all participants. One of the learnings from the study conducted earlier on in our project, was the skewness in the training data for gender. We then incorporated the gender of a subject as a feature to alleviate the issue and mindfully advertised our study to female volunteers to overcome this bias. In some cases, we also discovered that following the Hawthorne Effect (43), some participants were reluctant to report any higher fatigue throughout the task. Given a factory setting where higher or lower fatigue reports may draw administrative attention, users may be biased and stick to reporting neutral scores. Throughout this factory demonstration, we were able to obtain 75% of the total data collected due to sensor issues like perspiration, obstructed skin-to-sensor contact, etc. Adhesion can be affected by many factors such as sensor location, sweat, or contact with clothing and equipment. The conductive hydrogel adhesive used to secure the wearable sensors ensured skin-to-sensor contact, in most cases. However, excessive perspiration or equipment contact with sensor can remove the sensor from the body temporarily. This is disruptive towards the manufacturing working environment. To alleviate these potential issues, an additional layer of tape was applied on the sensors and skin to further secure their position. Most of the participants gave feedback that the wearable sensors were indeed unobtrusive and that such technology can enhance the quality of the working environment. This demonstration further supports our observation of

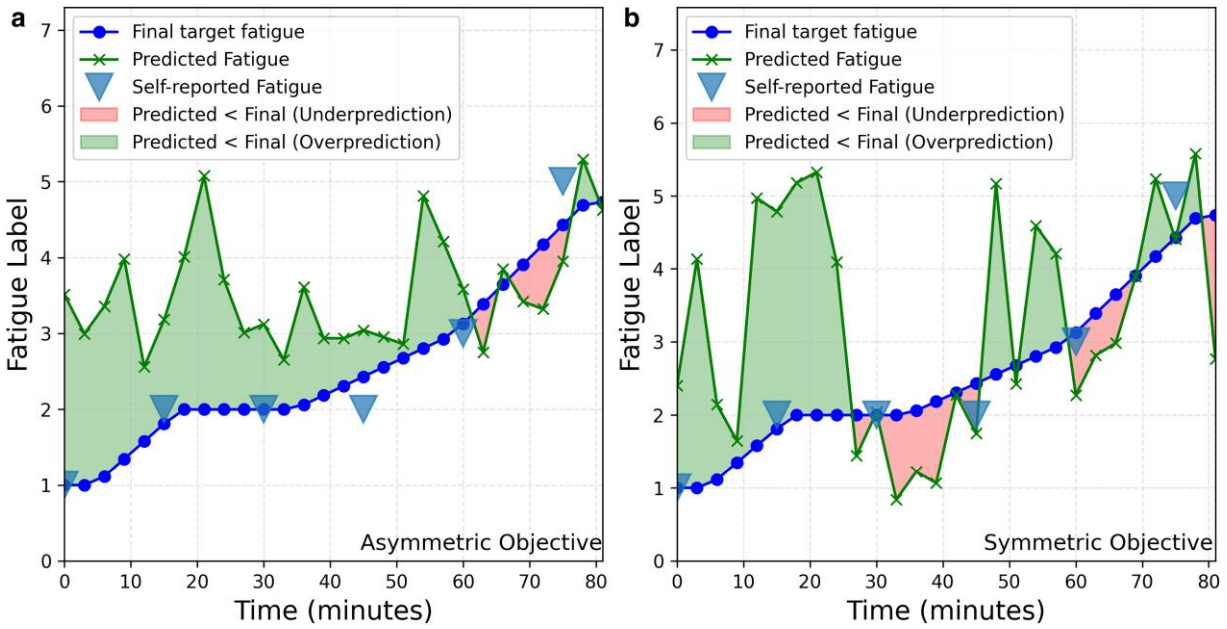

Fig. 4. Illustration comparing self-reported fatigue (blue markers) after each repetition with interpolated ground-truth fatigue (dotted line) used as the target for the predictive model. Predicted fatigue (cross-marked line) using a) an asymmetric linear-exponential (LINEX) objective with mean-absolute error (MAE) of 2.01, emphasizing operator-centric design by penalizing underpredictions (underprediction rate, UR = 0.18), and b) a symmetric mean-squared-error minimization objective resulting in MAE of 2.82 and UR of 0.46.

transferability on unseen (nonoverlapping) subjects across tasks as reported in Table 2.

User feedback and adoption

The two tasks selected use manual labor as the main power source. Measuring the level of worker fatigue can interfere with the activities of the workers. Workers may feel hindered by the equipment installed, resulting in bias due to the perceived inconvenience of workers. Hence it was imperative to get the views of the personnel that will use this technology in the future. Important attributes for the system to be usable would be comfort, compact form factor to avoid hindrance or interference with the task, data security and reliability, and finally usefulness of the fatigue sensing capability. To this end, all our participants were asked to complete a survey following the completion of tasks. The question prompts relate to (i) ease of wearing and removing the sensor, (ii) comfort on skin, (iii) hindrance to motion, and (iv) agreement to data tracking. Each question was answered with a score from 1 to 5, with 1 and 5 corresponding to complete agreement or disagreement, respectively, with the prompt. Eighty-six different responses across two tasks were collected from all Northwestern experiment participants and are presented in Fig. S4. Three responses were collected from the "in-the-wild" study and are presented in Fig. S5. The mean response for the questions relating to ease of application, comfort, and future use was 4 out of 5. The mean response for the question relating to hindrance was 2 out of 5, which is favorable (1 relates to no hindrance and 5 relates to maximum hindrance). Overall, a positive response was obtained from the Northwestern University participants, with a general consensus that the sensor system is unobtrusive and easy to use.

Discussion

Previous surveys across different countries indicate a high prevalence of fatigue among manufacturing workers, with potential consequences including increased injury risk, reduced production, and various short-term and long-term health issues. The improvement of working ergonomics and fatigue mitigation would greatly benefit manufacturing workers. Current challenges include the need for unobtrusive fatigue sensing methods and uncertainty regarding biomarkers. The adoption of new technologies for real-time fatigue prediction holds the potential to revolutionize manufacturing by optimizing work schedules and implementing adaptive work/rest cycles, addressing the issue of a lack of deterministic biomarkers.

Systematic large-scale studies on biomarkers for physical fatigue are rare. In this research, we focus on modeling fatigue as a continuous target and collect data in a more real-world manufacturing setting from 43 subjects for two tasks. We evaluate various on-body sensing modalities and study their influence on fatigue state.

Our first major finding is that, for true meaningful feedback about the operator's fatigue, we need to view the fatigue as a continuous variable. The past works of classifying a state as fatigued vs. nonfatigued are not very informative for taking preemptive safety measures. We showcase a regression-based formulation for fatigue-state and predict worker's fatigue on a scale of 0–10. Moreover, we drop any assumption of a worker being in a rest state (fatigue level 0) initially. This allows for a more practical approach to developing a closed-loop real-time fatigue prediction system.

Our second major finding is that modeling fatigue is very complex due to noisy self-reported labels. As we show quantitatively, it is also highly person- and task-specific. The nature of the task dictates the order of importance of various biomarkers. For example, for Task Composite which has more synchronized movements, PCA of feature vectors as well as Shapely scores of the trained model agree on the influence of locomotive features on fatigue state. Also, we show that testing a model on an entirely new individual whose data have not been used for training the model, has about 21% drop in the average performance. This is reasonably expected since the nonoverlapping users' data are out-of-distribution from the sample of training data. The key is to identify the inter-person variability causing this gap in performance.

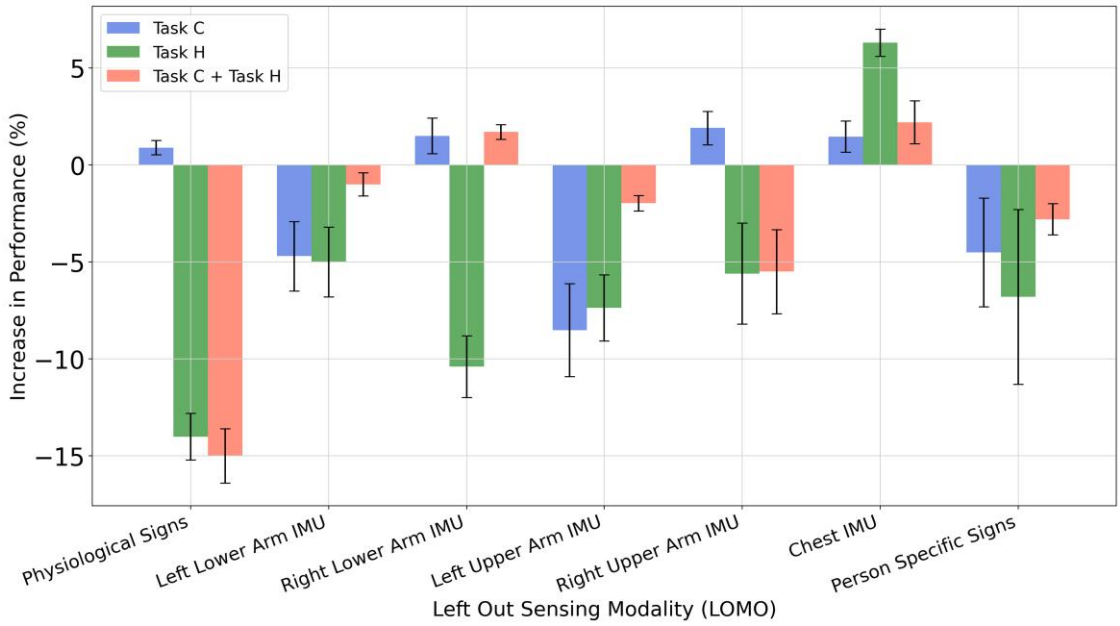

Fig. 5. Leave-one-modality-out analysis: a study to show the impact on model Performance when trained with one of the features left out (horizontal axis) in Task Composite (C), Task Harnessing (H), and a combination of Task Composite and Task Harnessing. The vertical axis represents the relative increase in performance for each task for their respective baseline mean-absolute-error (Task C: 2.45, Task H: 2.29, Task C + Task H: 2.41).

Our third finding is that the objective of the machine learning module needs to be more human-centric for such applications. It should penalize under-prediction (the operator is more fatigued than what the model predicts) more severely than over-prediction (the operator is less fatigued than what the model predicts). We propose the usage of an asymmetric loss function (Linear Exponential Loss [LINEX]) for optimizing our machine learning model which helps reduce under-prediction by approximately 50% compared to the traditional mean-squared-error loss function.

We are able to deploy a closed-loop system of continuously monitoring data from a worker using various wearable sensors to predict fatigue using a machine-learning-based inference engine and provide feedback using a visual dashboard. The visual dashboard in Fig. 1d shows the most informative sensing data to allow the users to understand the health trends that lead to the predicted fatigue state. We have effectively validated this system in real-world factory settings with employees, reinforcing our confidence in the technology's effectiveness. User feedback has consistently indicated a high level of acceptance and practical utility within the manufacturing industry. Although our overarching goal through this research is to ensure worker safety, mitigate risks, and empower operators through active feedback, we recognize the ethical and legal considerations associated with deploying such systems in real-world workplace environments. We are hopeful that ongoing technical advancements, including our efforts in predicting physical fatigue in manufacturing settings, will inspire constructive discussions about deployment.

Materials and methods

Study design

Monitoring fatigue is challenging due to the subjective nature of the perception of fatigue across various demographics. The study was designed to minimize these subjective differences to a maximum extent, to avoid biases in the predictions. The task protocol followed for data collection was consistent between the two tasks. All task steps and protocols were approved by the Northwestern University Institutional Review Board and subjects provided written informed consent. A total of six wearable sensors (one ANNE and five ADAM sensors) were used along with a vision system comprising two depth cameras (Intel Realsense D435 Depth Camera) and a regular HD webcam (Logitech C920x). The ANNE sensor was placed on the subject's left upper torso, below the collarbone. The ADAM sensors were located at five different locations: upper and lower arm (bicep and inner forearm) for both sides and upper torso as shown in Fig. 1a. Depth Camera 1 was set up in front of the workspace to capture the joint movements of the subject. Depth Camera 2 was used to capture data during moments of occlusion in Camera 1's data stream. The webcam was set on the ceiling with a bird eye's view to record the task. This stream was used to calculate task performance metrics (refer to Fig. S3). The field of view (FOV) for the camera system subcomponents remained the same for every task session to maintain data consistency.

All six wearable sensors were sterilized with alcohol swabs prior to application onto the subject. The soft, flexible wearable sensor patch was secured onto the subject with the help of a disposable double-sided hydrogel adhesive. Finally, tape was applied on top of each sensor to ensure skin contact throughout the task time. Subjects were asked to provide personal details such as age, weight (lbs.), and height (cm). Subjects were then asked to put on an ageing vest (weighted vest) along with 10 lbs. in wrist weights. The vest allows for weights up to 40 lbs. in increments of 2.5 lbs. Subjects were given the freedom to select a comfortable weight level but were encouraged to take a challenging weight for each repetition to ensure the development of fatigue signs. The tasks to perform (Task Composite and Task Harnessing) are repetitive and split into five repetitions. The task begins and ends with a 5 minute rest period. Including rest periods, there are 7 different data segments generated for each task (two rest periods + five repetitions). Subjects were asked to periodically (after each repetition and after rest periods) fill out a survey form with subjective questions regarding their state of

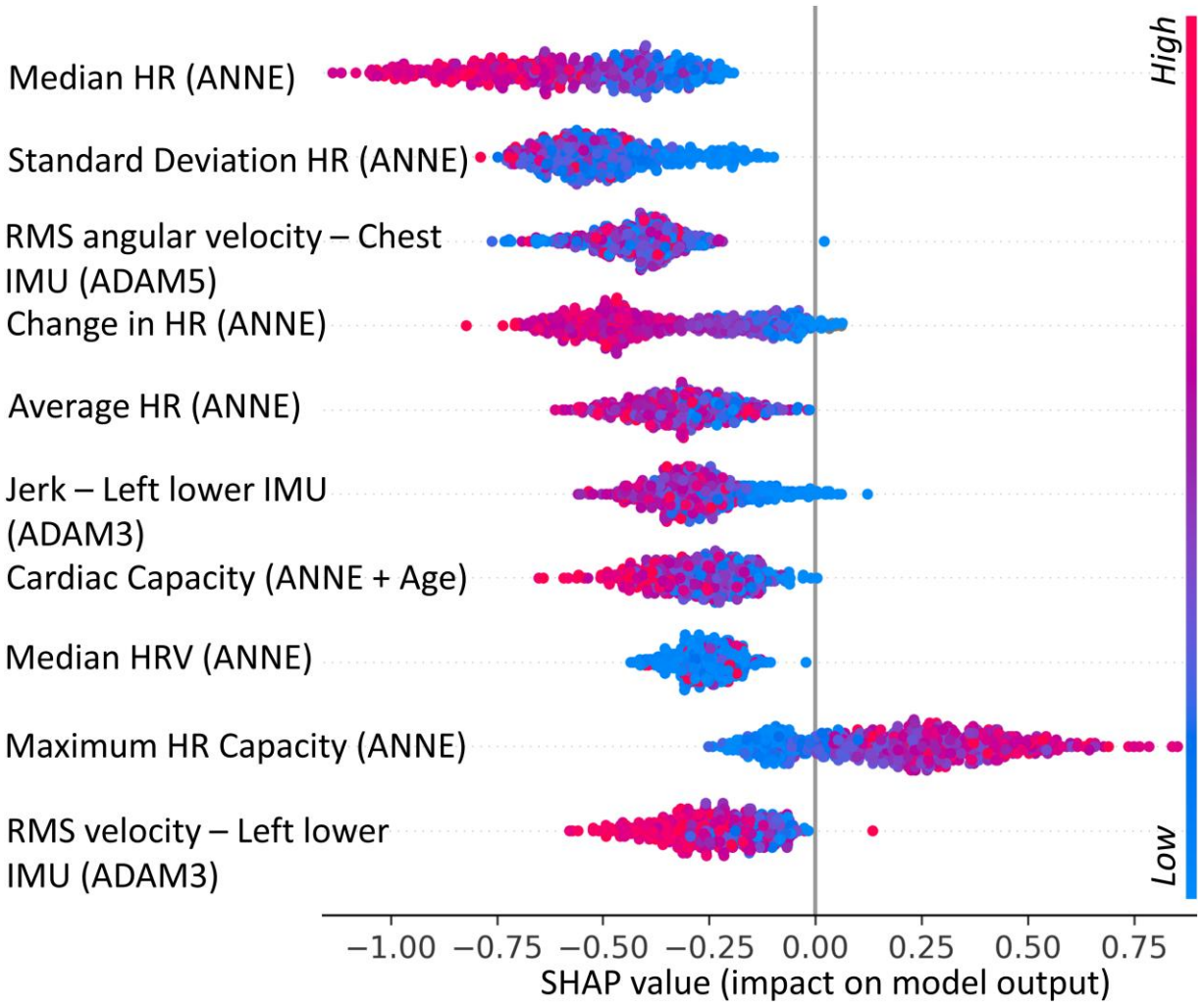

Fig. 6. Tree-based Shapley explainer for combination (Task Composite + Task Harnessing) tasks for all subjects.

fatigue. Of particular interest is the fatigue level selected by the subject after each period (repetition/rest). This value, ranging from 0 to 10 (0—no fatigue, 10—max fatigue) is used to quantify the subject's fatigue level to help with model training. All user inputs (age, weight, height, fatigue levels, survey questions, ageing vest weight) were recorded on an iPad tablet. After every data collection session, the sensors were disconnected and switched to data transmission mode by placing them onto the charging pads. The data were stored from each sensor in the form of raw files (.shrd). The raw files can be converted to .csv format with a file type converter.

Task protocol

We are motivated to use work-relevant manufacturing setups to simulate fatigue-inducing environments. Two experimental setups were designed to conduct the study. These tasks allow participants to get fatigued through physical exertion. An important consideration, in this case, would be the workload and time taken to complete the task. The format of the tasks is repetitive rather than sequential. This would allow us to compare data from similar segments of repeated motion at different points in time. Since we would like participants to try out both tasks designed, a task time of around 1 h is chosen. This ensures that some, if not all, participants show signs of fatigue towards the end of the task. To further accelerate exertion and fatiguing tendency, we used external weights to simulate an increased workload. Weighted vests (ageing suits) and wrist weights were used for this purpose. Ageing suits allow the subjects to feel the simulated fatigue in muscles. This is also helpful in mimicking a real environment, considering that a lot of subjects may be younger than the typical workforce that usually performs these tasks. No adverse events were reported during the course of this study.

Task Composite—Composite Layup

This setup simulates a typical composite sheet layup task, involving placing and smoothing out carbon fiber sheets on a complex mold geometry. The sheets will be placed according to a predefined layout. Twenty-four sheets of different sizes are placed on the mold, in order, and smoothed out. The smoothing motion is important, even though it makes no difference in the stickiness of the sheets in the mock setup (due to the absence of resin). In an actual setting, the sheets would need to be smoothed out onto a mandrel with resin, to make sure it sticks and takes the desired shape. It is important to try and mimic this motion to get an accurate representative task. Thin magnets have been provided to provide some bonding between the sheets. The colored lines, as shown in Task Composite image in Fig. 1b, are used for tracking performance metrics.

Task Harnessing—Wire Harnessing

Electrical wiring harnesses are critical components in many manufacturing products. Due to their complexity, they are one of the highest warranty items. Identification of root causes in the field can sometimes be very time-consuming and expensive, leading to customer dissatisfaction. Many internal standards have been developed over the years to improve the design and assembly of wiring harnesses. A significant amount of operator training has been done. However, there is a continuous turnover of factory workers, and often inexperienced operators could potentially install these complex assemblies. The installation of wiring harnesses can be complicated. Some harnesses can weigh as much as 30 lbs. and need to be assembled in difficult-to-reach areas on a machine. This setup is used to simulate the task of applying zip ties to a cable system, performed by workers in a manufacturing setting. A test bed is designed to accommodate the cable system. Participants are asked to apply zip ties to predefined zones. This task allows participants to mimic positions and postures taken by real manufacturing workers during a wire harnessing task. Nineteen different zones across a cable system are marked on the test bed. Subjects need to apply zip ties/cable ties at these zones, tighten them, and trim them. The dimensions of the test bed allow the worker to stay within range of the depth-sensing camera. Test bed images and dimensions can be referred to in Figs. S1 and S2.

Sensing framework

We use a multisensor framework to continuously monitor the physiological and locomotive signatures of the subjects throughout the task duration noninvasively. We primarily use a wearable subsystem of sensors on the body for understanding the fatigue biomarkers. We also have a vision subsystem that is designed to capture the skeletal pose estimations of the worker and not the raw video footage. This is used to quantify performance metrics for the tasks.

The standard for monitoring physiologic function in individuals (including heart rate, ECG, respiratory rate, temperature, etc.) requires constant, wired attachment to a power supply and operators, which can limit visibility and impair the ability to function normally in a manufacturing setting. To simulate an ideal manufacturing setting, the sensor system must be small and compact in form factor and must conform to the skin surface effectively to prevent hindrance and disturbances during operation. The wearable sensors were originally developed at Northwestern University and manufactured by Sibel Health. The ANNE chest sensor and ADAM sensor have been used in multiple studies (ANNE (32, 44, 45) and ADAM (33, 46–48)). The sensors have been cleared by the Food and Drug Administration (FDA).

ANNE chest sensor is a flexible wearable device placed on the torso of the subject. The sensor provides ECG, respiratory rate (RR), seismocardiography, body orientation, skin temperature, activity levels, and vocal biomarkers (crying patterns, cough index) (44). The ANNE chest sensor includes a biopotential analog front end (AFE) (MAX30001; Maxim Integrated), a high-frequency three-axis inertial measurement unit (IMU) (LSM6DSL; STMicroelectronics), and a clinical-grade thermometer (MAX30205; Maxim Integrated) (45). The device can operate continuously and wirelessly for up to 60 h, relying on wirelessly rechargeable lithium-polymer batteries (60 mAh). The data storage capacity is 4GB on the device. The electronic components are encapsulated with soft, flexible medical-grade polyorganosiloxane materials, and secured onto the skin using a thin, conductive hydrogel adhesive (KM 40A, Katecho). The ANNE sensor is placed on the subject's upper left torso during the experiments. ADAM sensor is a high-resolution, soft, flexible, wearable mechanoacoustic device. The sensor makes use of a flexible printed circuit board (fPCB; 25-μm-thick middle polyimide with double-sided 12-μm-thick rolled, annealed copper, AP7164R, DuPont) with serpentine conductive traces. Chip scale components include a high-bandwidth, inertial measurement unit (IMU) with a triaxial accelerometer (LSMDSL, STMicroelectronics) serving as the key sensing element, a Bluetooth Low Energy (BLE) system-on-a-chip (SoC) for control and wireless connectivity, on-device memory module for data storage, and a wirelessly rechargeable compact battery unit (47). The electronics are encapsulated by a thin, soft, elastomer membrane (Ecoflex, 00–30, smooth on, 300 μm) serving as a compliant, nonirritating interface when secured using a thin double-sided biomedical adhesive. For this study, the ADAM sensor is placed on the subject's bicep and inner forearm (both arms), as well as the upper torso.

The entire system is compatible with iOS devices for real-time streaming or on-sensor data storage synchronized with the cloud (HIPAA compliant/HITRUST certified web application). The myRA software application is available on iOS and enables multiple sensors to be automatically linked and time synchronized. The software further allows for both real-time streaming and data download features with raw data for further follow-on analysis.

Analysis framework

Data preparation

All the sensor data are time-synchronized and down-sampled to have a uniform sampling rate of 500 ms. As per the experiment protocol, the participants jump at the end of a repetition. This causes a spike in the ADAM sensors' (IMU) data which serves as the marker for the duration of the repetition with a simple thresholding technique. Figure 1b illustrates the experiment setup and a timeline of the task protocol. Additional information regarding the raw signals obtained from the sensors can be found in the Figs. S16 and S17.

Data denoising and filtering

The ECG measurements from the ANNE sensor are sensitive to ambulatory signals. We leverage the R-peak detection algorithm from Lee et al. (33) with slight modification in the filtering scheme (49, 50). We have used a Least Mean Square adaptive filter for noise cancellation followed by Band-pass filtering. The peak detector (33) extracts RR-time series and computes the temporal Heart Rate and Heart Rate Variability features. The motion-related signals measured from the ADAM sensors are filtered using a low pass filter with a cut-off of 15 Hz. Figure 1c shows the data preparation steps to convert the highly granular data to a tabular format to input into the machine learning model. These data are filtered based on the signal quality index (SQI) which is computed based on the correlation between the ECG signal and the accelerometer data and is a number between 0 and 1. Any segment with a cumulative score less than 0.5 is discarded to maintain the input data quality of the machine learning model. From the total of 43 participants' data, we have successfully processed 41 datasets for Task Composite and 36 datasets for Task Harnessing, following the SQI check.

Feature engineering

For every participant, we can effectively extract seven such repetitions for a given task as illustrated in Fig. 1b—initial rest segment, five repetitions of the task, and a final rest segment. The next step in the machine learning pipeline is the feature engineering of the input space in a more informative format. The denoised signal stream

Table 3. Summary of key features derived from the network of ANNE and ADAM sensors.

Signal category	Features
Physiological/Vital signs	Heart rate (HR), Heart rate variability (HRV), Skin temperature (statistically derived features—mean, standard deviation, skewness, kurtosis)
Ambulatory Features	Chest IMU (ADAM5), Right Upper Arm IMU (ADAM1), Left Upper Arm IMU (ADAM3), Right Lower Arm IMU (ADAM2), Left Lower Arm IMU (ADAM4) (jerk, velocity, range of angular motion, statistically derived features—mean, standard deviation, skewness, kurtosis)
Person-specific Features	Age, Gender, Cardiac Capacity, Weight, Height, Kinetic Expense
EMA	Perceived Fatigue Label, Feedback on usability of the sensor-subsystem

is now sliced into fixed-size windows. Figure S13 illustrates the trend of mean absolute error for different window lengths and the best performance is achieved with a window of 3 min.

Table 3 summarizes the features from all the sensors, and Table S2 contains the comprehensive list of all the derived features. In addition to traditional statistical features like mean, median, kurtosis, and skewness, we incorporate features influenced by individual attributes such as cardiac capacity and kinetic expense. We compute a dimensionless jerk using the formulation from Melendez-Calderon et al. (51). To quantify cardiac capacity, we employ a modified Karvonen (52) formula, setting it at 80% of the maximum heart rate, derived from the subject's initial resting heart rate. This is computed as

$$\text{Cardiac capacity} = 0.8((220 - \text{age}) - HR_{rest}), \tag{1}$$

where the mean heart rate approximates the resting heart rate (HR_{rest}) during the initial rest period. For kinetic expense estimation, we adopt a method inspired by Komaris et al. (53), utilizing the root mean square (RMS) velocity captured by ADAM5 as a proxy for the subject's center of mass velocity. This computation incorporates the subject's mass and that of any weighted vest worn during the activity. We compute,

$$\text{Kinetic expense} = 0.5(M_p + M_w)\sqrt{\frac{v_x^2 + v_y^2 + v_z^2}{3}}, \tag{2}$$

where M_p is the weight of the person, M_w is the weight of the vest donned during the experiment, and $v_i \forall\, i \in x, y, z$ is given as $v_i(t) = \int_0^t a_i(\tau)\, d\tau$ where t is the time-step of the accelerometer data a_i. All the features except the target fatigue labels are feature-scaled to a range of 0 to 10.

Data analysis

The reported fatigue label is a categorical variable in the range of 0–10. However, there is a physical implication of the degree of mispredictions in the fatigue label. We formulate our objective to minimize a mean-squared error between the true and predicted labels. This translates our task as a regression where our key objective is to map the input features to a continuous output target. We leverage this idea of continuous targets to interpolate the momentary fatigue score of the subjects to establish distinct fatigue labels per segment of the data. We interpolate fatigue scores across each repetition of the tasks based on the task duration using the equation $\text{fatigue}_{new} = \text{fatigue}_{initial} + (\text{fatigue}_{final} - \text{fatigue}_{initial})/\text{RepDuration}$. We choose a 3-min window for fatigue prediction with continuous fatigue targets between 0 and 10. Previously described feature engineering techniques are key to the design of machine-learning methods in data-poor settings. Using raw physiological and ambulatory data in a sequential capacity (54) did not provide promising results. From our preliminary analysis of limited datasets, gradient-boosted trees implemented using the eXtreme Gradient Boosting (XGBoost) (55) have outperformed shallow neural networks and that is our choice of design currently for regressors. Previous works have also supported the superior performance of tree-based models on tabular data (56) and specifically low-resource stress prediction tasks (57). Our motivation for the asymmetric loss function is to train a more operator-centric model that penalizes under-prediction more severely than over-predictions. From our results, we also observe that such a choice has a regularizing effect and compensates for the lack of data in higher fatigue regions. We use the Linear-Exponential (LINEX) loss function for instilling this asymmetric behavior to our objective as shown in Eq. 3. The degree of penalization is dictated by the hyperparameter a as shown in Fig. S15.

$$E \longrightarrow 2/a^2[e^{a\cdot E} - a \cdot E - 1],\ a < 0 \tag{3}$$

Training Scheme

We conduct our training and evaluations under the following key settings:

1. Overlapping and nonoverlapping users
2. Cross-tasks and task-specific
3. Subset of features

Under all these scenarios, we conduct a grid-search-based (58) hyper-parameter optimization for the regressor with a five-fold cross-validation. The training and testing were conducted on an Intel(R) Xeon(R) Gold 6,130 CPU at 2.10 GHz. Each experiment is run for three seeds (2711, 2712, 2713), and the mean and standard deviation statistics are reported to account for performance variability. For overlapping users, we use an 8:2 ratio for training and testing, respectively. For nonoverlapping users, we hold out data from 8 randomly selected participants from the pool of 43 participants. XGBoost is a lightweight model that supports model explainability. We leverage this property to understand the feature importance assigned by the model in making its prediction.

Evaluation metrics

To evaluate the performance of the models, we use Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Underprediction Rate (UR) as given below,

$$\begin{aligned} \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \\ \text{UR} &= \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 1 & \text{if } \hat{y}_i - y_i < 0, \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

where y_i is the ground-truth of fatigue score and $\hat{y}_i$ is the predicted fatigue score for the ith segment and n denotes the total number of segments.

Visualization dashboard

Due to the use of a gradient-boosted trees model, the model size is small and allows for near-real-time inference. The inference results and the feature-extraction scripts are used to translate data into a format acceptable to Google Data Studio. We design a dashboard using this backbone to allow participants to interact with their data and model predictions and collect their feedback. A mock-up dashboard is demonstrated in Fig. 1d.

Acknowledgments

The authors thank Nabil Alshurafa, Olivia Botonis, Arun Jayaraman, and Devashri Naik for their support and input during the pilot studies and the development process of this project.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This research was supported by MxD (Manufacturing × Digital Institute) under project numbers 19-13-05 and 22-06-02. This project was completed under the Technology Investment Agreement W15QKN-19-3-0003, between Army Contracting Command—New Jersey and MxD. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Army.

Data Availability

The deidentified raw dataset and all necessary meta-data are available at: https://zenodo.org/records/12788571. Our codebase for reproducing the main analyses can be found at: https://github.com/payalmohapatra/WorkerFatigue.git.

Ethics Statement

This study was conducted in full compliance with ethical standards and guidelines for human subjects research. The research protocol was approved by the Institutional Review Board of Northwestern University (approval number STU0021461) and received endorsement from the US Army Human Research Protections Office (reference number MXD191305). All participants provided informed consent before their involvement in the study.

References

1 Ricci JA, Chee E, Lorandeau AL, Berger J. 2007. Fatigue in the US workforce: prevalence and implications for lost productive work time. *J Occup Environ Med*. 49(1):1–10.
2 Yung M, Bigelow PL, Hastings DM, Wells RP. 2014. Detecting within- and between-day manifestations of neuromuscular fatigue at work: an exploratory study. *Ergonomics*. 57(10):1562–1573.
3 Loriol M. 2017. A sociological stance on fatigue and tiredness: social inequalities, norms and representations. *Neurophysiol Clin*. 47(2):87–94.
4 Kajimoto O. 2008. Development of a method of evaluation of fatigue and its economic impacts. In: Watanabe Y, Evengård B, Natelson BH, Jason LA, Kuratsune H, editors. *Fatigue Science for Human Health*. Tokyo: Springer. p. 33–46.
5 Evengård B. 2008. Fatigue: epidemiology and social/industrial aspects. In: Watanabe Y, Evengård B, Natelson BH, Jason LA, Kuratsune H, editors. *Fatigue Science for Human Health*. Tokyo: Spinger. p. 17–31.
6 Richter K, Acker J, Adam S, Niklewski G. 2016. Prevention of fatigue and insomnia in shift workers–a review of non-pharmacological measures. *EPMA J*. 7(1):1–11.
7 Björklund M, Crenshaw AG, Djupsjöbacka M, Johansson H. 2000. Position sense acuity is diminished following repetitive low-intensity work to fatigue in a simulated occupational setting. *Eur J Appl Physiol*. 81(5):361–367.
8 Côté JN, Raymond D, Mathieu PA, Feldman AG, Levin MF. 2005. Differences in multi-joint kinematic patterns of repetitive hammering in healthy, fatigued and shoulder-injured individuals. *Clin Biomech*. 20(6):581–590.
9 Iridiastadi H, Nussbaum MA. 2006. Muscle fatigue and endurance during repetitive intermittent static efforts: development of prediction models. *Ergonomics*. 49(4):344–360.
10 Fukuda K, *et al*. 1994. The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Ann Intern Med*. 121(12):953–959.
11 Jason LA, Benton MC, Valentine L, Johnson A, Torres-Harding S. 2008. The economic impact of ME/CFS: individual and societal costs. *Dyn Med*. 7(1):6.
12 Papoutsakis K, *et al*. 2022. Detection of physical strain and fatigue in industrial environments using visual and non-visual low-cost sensors. *Technologies*. 10(2):42.
13 Li L, Xu X. 2019. A deep learning-based RULA method for working posture assessment. *Proc Hum Factors Ergon Soc Annu Meet*. 63(1):1090–1094.
14 Rizkya I, Syahputri K, Sari R, Anizar A, Siregar I. 2018. Evaluation of work posture and quantification of fatigue by rapid entire body assessment (REBA). In: IOP Conference Series: Materials Science and Engineering. Vol. 309. IOP Publishing. p. 012051.
15 Villalobos A, Cawley A. 2022. Prediction of slaughterhouse workers' RULA scores and knife edge using low-cost inertial measurement sensor units and machine learning algorithms. *Appl Ergon*. 98:103556.
16 Borg GAV. 1982. Psychophysical bases of perceived exertion. *Med Sci Sports Exerc*. 14(5):377–381.
17 King ZD, *et al*. 2019. Micro-stress EMA: a passive sensing framework for detecting in-the-wild stress in pregnant mothers. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. 3(3):1–22.
18 Cohen S, Kamarck T, Mermelstein R. 1983. A global measure of perceived stress. *J Health Soc Behav*. 24(4):385–396.
19 Birkett MA. 2011. The trier social stress test protocol for inducing psychological stress. *J Vis Exp*. (56):e3238.
20 Maman ZS, Yazdi MAA, Cavuoto LA, Megahed FM. 2017. A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. *Appl Ergon*. 65:515–529.
21 Wong DP-L, Chung JW-Y, Chan AP-C, Wong FK-W, Yi W. 2014. Comparing the physiological and perceptual responses of construction workers (bar benders and bar fixers) in a hot environment. *Appl Ergon*. 45(6):1705–1711.
22 De Beéck TO, Meert W, Schütte K, Vanwanseele B, Davis J. 2018. Fatigue prediction in outdoor runners via machine learning and sensor fusion. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM. p. 606–615.
23 Hajifar S, *et al*. 2021. A forecasting framework for predicting perceived fatigue: using time series methods to forecast ratings of perceived exertion with features from wearable sensors. *Appl Ergon*. 90:103262.

24 Baghdadi A, *et al.* 2021. Monitoring worker fatigue using wearable devices: a case study to detect changes in gait parameters. *J Qual Technol*. 53(1):47–71.
25 Hellhammer DH, Wüst S, Kudielka BM. 2009. Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology*. 34(2): 163–171.
26 Kirschbaum C, *et al.* 1995. Persistent high cortisol responses to repeated psychological stress in a subpopulation of healthy men. *Psychosom Med*. 57(5):468–474.
27 Abuwarda Z, Hegazy T, Oetomo A, Morita PP. 2022. Using wearables to monitor and mitigate workers' fatigue. In: Proceedings of the Canadian Society of Civil Engineering Annual Conference 2021: CSCE21 Construction Track Vol. 2. Springer. p. 587–597.
28 Baghdadi A, Megahed FM, Esfahani ET, Cavuoto LA. 2018. A machine learning approach to detect changes in gait parameters following a fatiguing occupational task. *Ergonomics*. 61(8): 1116–1129.
29 Zhang L, Diraneyya MM, Ryu J, Haas C, Abdel-Rahman E. 2019. Automated monitoring of physical fatigue using jerk. In: Al-Hussein M, editor. Proceedings of the 36th International Symposium on Automation and Robotics in Construction (ISARC). Banff, Canada: International Association for Automation and Robotics in Construction (IAARC). p. 989–997.
30 Lee Y-J, Wei M-Y, Chen Y-J. 2022. Multiple inertial measurement unit combination and location for recognizing general, fatigue, and simulated-fatigue gait. *Gait Posture*. 96:330–337.
31 Donati M, Olivelli M, Giovannini R, Fanucci L. 2023. ECG-based stress detection and productivity factors monitoring: the real-time production factory system. *Sensors*. 23(12):5502.
32 Chung HU, *et al.* 2020. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. *Nat Med*. 26(3):418–429.
33 Lee KH, *et al.* 2020. Mechano-acoustic sensing of physiological processes and body motions via a soft wireless device placed at the suprasternal notch. *Nat Biomed Eng*. 4(2):148–158.
34 Williams N. 2017. The borg rating of perceived exertion (RPE) scale. *Occup Med (Chicago, IL)*. 67(5):404–405.
35 Tempelaar D, Rienties B, Nguyen Q. 2020. Subjective data, objective data and the role of bias in predictive modelling: lessons from a dispositional learning analytics application. *PLoS One*. 15(6): e0233977.
36 Baker RS, Hawn A. 2021. Algorithmic bias in education. *Int J Artif Intell Educ*. 32:1052–1092.
37 Song F, Guo Z, Mei D. 2010. Feature selection using principal component analysis. In: 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization; Yichang, China. Vol. 1. IEEE Computer Society. p. 27–30.
38 Spearman C. 1987. The proof and measurement of association between two things. *Am J Psychol*. 100(3/4):441–471.
39 Kim H-G, Cheon E-J, Bai D-S, Lee YH, Koo B-H. 2018. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig*. 15(3):235–245.
40 Taelman J, Vandeput S, Spaepen A, Van Huffel S. 2009. Influence of mental stress on heart rate and heart rate variability. In: 4th European Conference of the International Federation for Medical and Biological Engineering: ECIFMBE 2008 23-27 November 2008 Antwerp. Belgium: Springer. p. 1366-1369.
41 Lundberg SM, Lee S-I. 2017. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 30:4768–4777.
42 Lu L, Megahed FM, Sesek RF, Cavuoto LA. 2017. A survey of the prevalence of fatigue, its precursors and individual coping mechanisms among US manufacturing workers. *Appl Ergon*. 65: 139–151.
43 Adair JG. 1984. The hawthorne effect: a reconsideration of the methodological artifact. *J Appl Psychol*. 69(2):334–345.
44 Liu C, *et al.* 2021. Wireless, skin-interfaced devices for pediatric critical care: application to continuous, noninvasive blood pressure monitoring. *Adv Healthc Mater*. 10(17):e2100383.
45 Ryu D, *et al.* 2021. Comprehensive pregnancy monitoring with a network of wireless, soft, and flexible sensors in high- and low-resource health settings. *Proc Natl Acad Sci U S A*. 118(20): e2100466118.
46 Jeong H, *et al.* 2021. Differential cardiopulmonary monitoring system for artifact-canceled physiological tracking of athletes, workers, and COVID-19 patients. *Sci Adv*. 7(20):eabg3092.
47 Ni X, *et al.* 2021. Automated, multiparametric monitoring of respiratory biomarkers and vital signs in clinical and home settings for COVID-19 patients. *Proc Natl Acad Sci U S A*. 118(19): e2026610118.
48 Lonini L, *et al.* 2021. Rapid screening of physiological changes associated with COVID-19 using soft-wearables and structured activities: a pilot study. *IEEE J Transl Eng Health Med*. 9:1–11.
49 Mohapatra P, Premkumar PS, Sivaprakasam M. 2018. A yellow–orange wavelength-based short-term heart rate variability measurement scheme for wrist-based wearables. *IEEE Trans Instrum Meas*. 67(5):1091–1101.
50 Mohapatra P, Preejith SP, Sivaprakasam M. 2017. A novel sensor for wrist based optical heart rate monitor. In: 2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE. p. 1–6.
51 Melendez-Calderon A, Shirota C, Balasubramanian S. 2021. Estimating movement smoothness from inertial measurement units. *Front Bioeng Biotechnol*. 8:558771.
52 Karvonen MJ. 1957. The effects of training on heart rate; a longitudinal study. *Ann Med Exp Biol Fenn*. 35:307–315.
53 Komaris D-S, Tarfali G, O'Flynn B, Tedesco S. 2022. Unsupervised IMU-based evaluation of at-home exercise programmes: a feasibility study. *BMC Sports Sci Med Rehabil*. 14(1):28.
54 Mohapatra P, Pandey A, Keten S, Chen W, Zhu Q. 2023. Person identification with wearable sensing using missing feature encoding and multi-stage modality fusion. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. p. 1–2.
55 Chen T, Guestrin C. 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery. p. 785–794.
56 Grinsztajn L, Oyallon E, Varoquaux G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Adv Neural Inf Process Syst*. 35:507–520.
57 Lewis RA, *et al.* 2023. Mixed effects random forests for personalised predictions of clinical depression severity. *arXiv*, arXiv:2301.09815. https://doi.org/10.48550/arXiv.2301.09815, preprint: not peer reviewed.
58 Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. *J Mach Learn Res*. 13(2):218–305.