Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

A data analytic end-to-end framework for the automated quantification of ergonomic risk factors across multiple tasks using a single wearable sensor

Saeb Ragani Lamooki[a], Sahand Hajifar[b], Jiyeon Kang[c], Hongyue Sun[b], Fadel M. Megahed[d], Lora A. Cavuoto[b,*]

[a] Department of Mechanical Engineering, University at Buffalo, Buffalo, NY, 14260, USA
[b] Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, 14260, USA
[c] Department of Mechanical and Aerospace Engineering, University at Buffalo, Buffalo, NY, 14260, USA
[d] Farmer School of Business, Miami University, Oxford, OH, 45056, USA

ARTICLE INFO

Keywords:
Human performance modeling
Ergonomic risk assessment
End-to-end learning
Change point detection
Wavelets
Wearable sensors

ABSTRACT

Existing ergonomic risk assessment tools require monitoring of multiple risk factors. To eliminate the direct observation, we investigated the effectiveness of an end-to-end framework that works with the data from a single wearable sensor. The framework is used to identify the performed task as the major contextual risk factor, and then estimate the task duration and number of repetitions as two main indicators of task intensity. For evaluation of the framework, we recruited 37 participants to complete 10 simulated work tasks in a laboratory setting. In testing, we achieved an average accuracy of 92% for task identification, 7.3% error in estimation of task duration, and 7.1% error for counting the number of task repetitions. Moreover, we showed the utility of the framework outputs in two ergonomic tools to estimate the risk of injury. Overall, we indicated the feasibility of using data from wearable sensors to automate the ergonomic risk assessment in workplaces.

1. Introduction

Based on the 2015 *National Health Interview Survey*, one in two U.S. adults reported a musculoskeletal medical condition (Joint Initiative et al., 2020a). The total direct and indirect costs of musculoskeletal disorders (MSDs) was estimated to be $980.1 billion per year in 2012–2014, a 5.76% share of the gross domestic product (GDP) (Joint Initiative et al., 2020b). More alarmingly, the costs of MSDs have been rising steadily. There are several well-known risk factors associated with MSDs, including tasks involving high repetitions, high force demands, awkward postures, and long duration of physical exertion (Putz-Anderson et al., 1997; Da Costa and Vieira, 2010; Gallagher and Heberger, 2013). Other personal risk factors include co-morbid diseases and psychosocial/physiological conditions (National Research Council, 2001). In this article, we will focus on physical risk factors (repetitions, duration, and force demands), since personal risk factors are less controllable from a job design perspective.

Existing ergonomic risk assessment tools require the accurate collection and measurement of one or more of the physical risk factors. Examples include the NIOSH Lifting Equation (Waters et al., 1994), Rapid Entire Body Assessment (REBA) (Hignett and McAtamney, 2000), Rapid Upper Limb Assessment (RULA) (McAtamney and Corlett, 1993), Ovako Working Posture Assessment System (OWAS) (Karhu et al., 1977) as well as the more recent fatigue-failure based Lifting Fatigue Failure Tool (LiFFT) (Gallagher et al., 2017) and The Shoulder Tool (Bani Hani et al., 2021). Observational approaches are heavily used by professional ergonomists in industry and their use has increased from 2005 to 2018 (Lowe et al., 2019). These approaches typically rely on visual inspection of occupational tasks by a trained professional who would measure/estimate the values for the different physical factors.

There are multiple reasons why the reliance on trained professionals can be somewhat limiting in practice. First, the reliance on a human observer in the field is costly due to labor and time considerations (Takala et al., 2010). Hence, the number of observers is much smaller than the jobs performed, which often results in restricting risk assessments to job design in repetitive environments and/or periodic audits. Second, the variability in implementing different ergonomic risk tools is typically ignored in practice. From a measurement systems analysis

[*] Corresponding author.
E-mail addresses: saebraga@buffalo.edu (S.R. Lamooki), sahandha@buffalo.edu (S. Hajifar), jiyeonk@buffalo.edu (J. Kang), hongyues@buffalo.edu (H. Sun), fmegahed@miamioh.edu (F.M. Megahed), loracavu@buffalo.edu (L.A. Cavuoto).

https://doi.org/10.1016/j.apergo.2022.103732
Received 14 October 2021; Received in revised form 18 February 2022; Accepted 24 February 2022
Available online 12 March 2022
0003-6870/

perspective, a repeatability and reproducibility (R&R) analysis (Burdick et al., 2003) should be performed to assess an organization's capability of utilizing a specific risk assessment tool. The R&R analysis is based on an analysis of variance (ANOVA) random effects model that captures the tool's: (a) *repeatability*, whether an ergonomic professional's repeated assessments of a given task result in similar risk scores; and (b) *reproducibility*, whether different observers' assessments/scores are consistent. Third, many occupational environments are changing and exhibit an "increasing heterogeneity of workers, work, and the workplace" (National Research Council, 1999, p. 3). Thus, there is a need for frequent, possibly continuous, assessment of ergonomic risk to capture changes in work and workplaces. Fourth, fatigue can cause an individual to perform tasks differently, with less control over time (Cortes et al., 2014), and this is often not observable by the human eye (Baghdadi et al., 2021).

Wearable sensors have emerged as a solution to overcome the aforementioned limitations in observational approaches to ergonomic risk assessments. They are portable, lightweight, inexpensive, non-invasive and have battery lives and bandwidth that allow for continuous monitoring over the course of a work shift (Cavuoto and Megahed, 2017; Camomilla et al., 2018; Lim and D'Souza, 2020; Stefana et al., 2021). In the review of Lim and D'Souza [21, p. 8], the authors concluded that "these studies demonstrated advantages of using direct inertial sensing measurements over observation- and questionnaire-based methods which are time and labor intensive and more susceptible to measurement bias". Furthermore, their portability and the relative efficiencies in analyzing their data streams have made them a preferred alternative to camera-based motion tracking systems in practice (Scott and Browning, 2016).

Current applications of wearables for biomechanical exposure assessment span modeling, sensing, analysis, assessment and/or intervention (MSAAI) (Lim and D'Souza, 2020). In our estimation, there are three major issues that need to be addressed in order to accelerate their adoption in the field. First, inertial sensors are designed to repeatedly capture acceleration/angular velocity profiles over time. However, they tend to lack context of the performed tasks. For example, Lim and D'Souza [21, p. 12] stated "most studies that examined biomechanical exposures by task (25 out of 30, 83.3%) incorporated additional methods (e.g., direct observations, self-reported measures) to identify task type and duration (i.e., start and end of specific tasks) since such information is not readily apparent from inertial sensor data alone." Note that the remaining five studies used multiple sensors (up to 17 different inertial measurement units (IMUs) in (Kim and Nussbaum, 2014)) and focused on postural analysis. Second, once the task is identified, the challenge is to extract the relevant biomechanical features, including the duration and number of repetitions for the task performed. Most studies reviewed in (Lim and D'Souza, 2020; Stefana et al., 2021) have focused on counting based on postural thresholds, e.g., the number of shoulder elevations beyond 45°. While posture is an important risk factor, repetition without load poses a different injury risk to the tissue (Gallagher and Heberger, 2013). The task context allows for more accurate determination of loaded repetitions versus basic arm movements. For example, a worker may be stretching, but a posture-based approach would likely identify such an activity as a task repetition. Third, reproducing the results of the published literature is difficult in practice due to the: (a) relatively large number of sensors used (median of 3 and up to 17) (Lim and D'Souza, 2020); and (b) lack of end-to-end analysis, covering everything from data acquisition and pre-processing to model training, validation and deployment.

The overarching objective of this paper is to develop an end-to-end data-centered framework for ergonomic risk assessment based on a single accelerometer, placed on the wrist, to maximize wearability/user comfort (Porta et al., 2021), and capitalize on the results from Mokhlespour Esfahani and Nussbaum (Mokhlespour Esfahani and Nussbaum, 2018) who reported that the wrist is a preferred location for placement of inertial sensors. Capitalizing on insights from the *CRoss Industry Standard Process for Data Mining* (CRISP-DM) (Wirth et al., 2000), our data-centered framework covers the steps from problem definition and sensor selection/placement up to modeling the different physical risk factors and utilizing such information to compute and display an appropriate ergonomic risk score. To accomplish this objective, we have examined the following research questions:

(1) How can we automatically divide the time series of acceleration signals, obtained from a wrist-worn accelerometer, into different segments? The goal is to examine how to optimize the search for possible change points/segment boundaries to capture task transitions from the wrist acceleration data.
(2) Within each segment, can we determine the correct task label for the segment? Here, we attempt to label/classify the unlabeled segments using supervised machine learning techniques.
(3) How accurately can we determine the total duration of each identified/performed task?
(4) How accurately can we determine the number of repetitions for repetitive tasks, such as the number of lifts or the number of hand pulls in a hoisting task.
(5) How can we fuse the physical ergonomic risk factors (i.e., task type, duration, repetition) from the accelerometer with load and task-design information to systematically estimate the probability of a workplace injury?

The remainder of the paper is organized as follows. Our end-to-end, conceptual framework for utilizing a single accelerometer for ergonomic risk assessment is presented in Section 2. In Section 3, we provide the technical methods needed to implement the conceptual framework. The results of the application of the methodology to the experimental study are provided in Section 4. Our conclusions and recommendations for future research directions are presented in Section 5.

2. Conceptual framework and related literature

2.1. An overview of the proposed end-to-end framework

In Fig. 1, we present an overview of our proposed end-to-end framework for ergonomic risk assessment using a single accelerometer. The framework extends the MSAAI framework of (Lim and D'Souza, 2020) by: (a) explicitly highlighting the importance of clearly defining the ergonomic problem of interest to ensure that the goals and requirements are distinctly communicated; (b) dividing the modeling stage into planning (in our first and second phases) and feature engineering (in our third phase); and (c) expanding on the steps in the analysis phase to include change point detection, task identification (activity recognition), and estimation of task duration/repetition count.

2.2. State-of-the-art pertaining to the phases of our conceptual framework

In phase 1, we define the assessment objective and scope of the ergonomic project in terms of capturing the work domain characteristics, identifying the jobs/tasks to be analyzed, the loading mechanisms for the tasks, the major body parts impacted by the tasks, and the fatigue indicators/outcome measures. Once the scope and objective(s) of the project are defined, the sensor type and its location on the body are to be determined in phase 2. The choice of sensor, its location, and usage for ergonomic assessment have been extensively discussed in the reviews of Lim and D'Souza (Lim and D'Souza, 2020), Stefana et al. (2021), and Ranavolo et al. (2018). Based on their findings and the discussion in Section 1, in this paper, we examine the use of a single accelerometer placed on the dominant wrist for ergonomic risk assessment.

In phase 3, the sensor signal is pre-processed prior to task analysis. The pre-processing of the acceleration signals commonly involves the application of low-pass filters to remove noise (Maman et al., 2020; Hajifar et al., 2021a) and/or feature engineering to improve the model

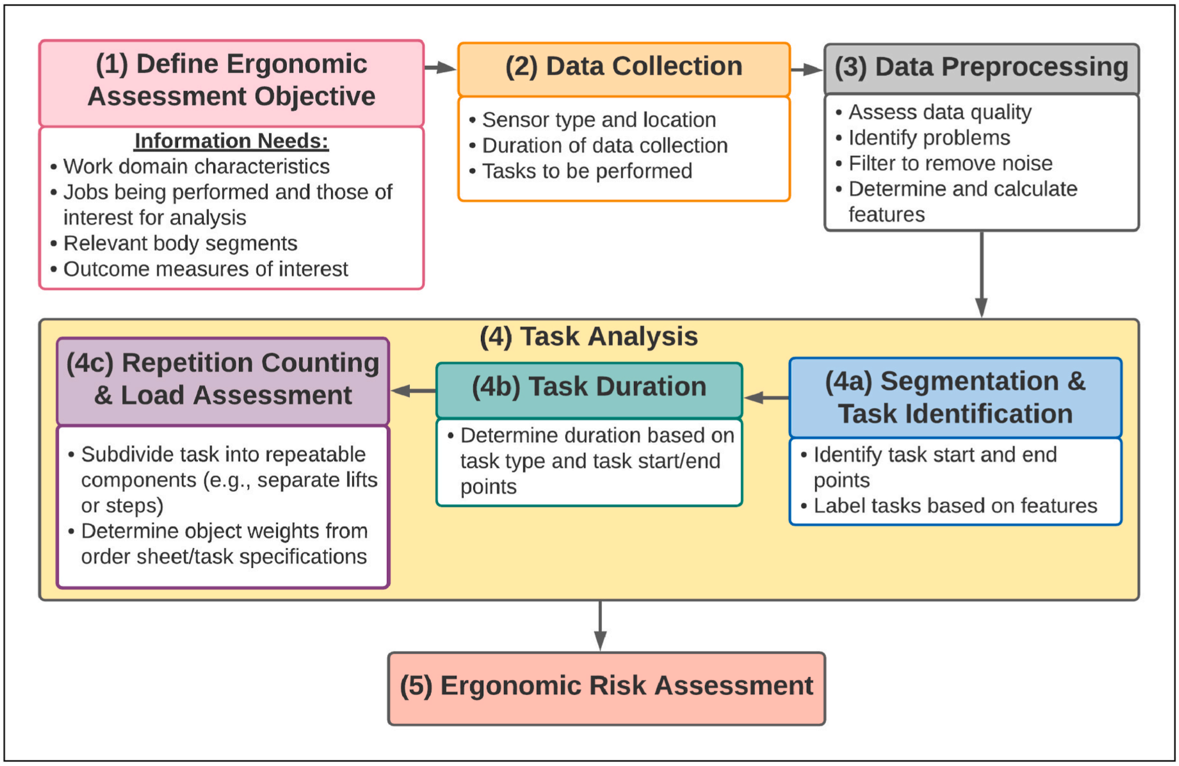
(1) Define Ergonomic Assessment Objective
Information Needs:
• Work domain characteristics
• Jobs being performed and those of interest for analysis
• Relevant body segments
• Outcome measures of interest
(2) Data Collection
• Sensor type and location
• Duration of data collection
• Tasks to be performed
(3) Data Preprocessing
• Assess data quality
• Identify problems
• Filter to remove noise
• Determine and calculate features
(4) Task Analysis
(4c) Repetition Counting & Load Assessment
• Subdivide task into repeatable components (e.g., separate lifts or steps)
• Determine object weights from order sheet/task specifications
(4b) Task Duration
• Determine duration based on task type and task start/end points
(4a) Segmentation & Task Identification
• Identify task start and end points
• Label tasks based on features
(5) Ergonomic Risk Assessment

Fig. 1. A conceptual framework for data-driven ergonomic risk assessment.

performance (Lim and D'Souza, 2020; Tsao et al., 2018; Maman et al., 2020; Hajifar et al., 2021b; Baghdadi et al., 2018; Maman et al., 2017). The choice of pre-processing technique should also account for how the analysis will be performed (i.e., offline vs real-time). For real-time applications, the computational and communication complexity becomes an important factor in determining the suitability of a given pre-processing technique or recipe.

In the 4th phase, the literature is disjointed and there are a limited number of papers that have examined all 3 steps of 4a–4c in Fig. 1. While many of the classification models used for ergonomic applications had good predictive performance (Lim and D'Souza, 2020; Stefana et al., 2021), they may be sub-optimal for repetition counting. Per Martindale et al. (Martindale et al., 2021, p. 250),

> Many of these tasks require not only the classification of activities, but also an analysis of the cyclic nature of them … Segmentation boundaries are important for the analysis of gait and activities in precise applications … such a task can be interpreted as a sequence-to-sequence problem, rather than the typical classification problem.

Hence, the reliance on classification models alone can result in inaccurate estimation of task duration, as detection of the start and end of a given task depend on the choice of window length when a windowing approach is used. Moreover, task repetition counting from wearable sensors is not widely investigated in ergonomic applications. In the following paragraphs we highlight some of the most relevant literature for each of the steps in phase 4.

There are three general approaches for segmentation/change point detection in signals from wearable sensors: (a) hidden Markov models (Panahandeh et al., 2013; Martindale et al., 2017); (b) change-point detection models for detecting statistically significant changes in the inertial time series, such as the Greedy Gaussian Segmentation (GGS) approach Li et al. (2019); and (c) deep learning techniques suitable for time-series data [e.g., see 35]. In this paper, we will utilize the GGS approach since it is computationally efficient (Hallac et al., 2019), has an open-source implementation (Hallac et al., Boyd), and allows us to capture change-points appearing in different sensor channels for different task transitions.

Various statistical and machine learning methods have been used for task identification with applications in ergonomics (Tsao et al., 2018; Lim and D'Souza, 2020; Stefana et al., 2021). Note that the pre-processing required for those models heavily depends on the type of models used. For example, statistical and traditional machine learning models often require the extraction/computation of statistical/kinematic features from the signals prior to the model building and deployment. On the other hand, neural networks can capitalize on their structure for feature extraction in the earlier layers. For example, neural networks have successfully been used on both raw inertial signals (Zeng et al., 2014; Ronao and Cho, 2016) and time-frequency domain features from short-time Fourier transform (Ravi et al., 2016) in activity recognition applications. The use of feature extraction techniques, such as short-time Fourier transform or Continuous Wavelet Transform (CWT), to extract time-frequency features allows for the reduction of the number of layers in neural network architectures. Application of CWT for feature extraction has been shown to improve the performance of Convolutional Neural Networks (CNN) in detecting falls, seizures, and arrhythmia from physiological signals (inertial, electroencephalogram (EEG), and electrocardiogram (ECG)) (Yhdego et al., 2019; Mao et al., 1456; Wang et al., 2021). Once the data is segmented and the segments are labeled/annotated, the task duration is obtained from the segment length (segment start and end).

Applications of task repetition counting include both sport/exercise (Chang et al., 2007; Pernek et al., 2013; Choi et al., 2013; Mortazavi et al., 2014) and industrial (Radwin and Lin, 1993; Douphrate et al., 2012; Álvarez et al., 2016; Peppoloni et al., 2016; Park and Kim, 2017) settings. In sport applications, the pace/frequency of the tasks can be controlled (e.g., walking on a treadmill), the movement patterns can be limited/restricted (e.g., weightlifting), and the subjects are committed to follow the same movement patterns to engage specific muscle groups. These characteristics are often absent in industrial applications, where the tasks are more complex and the workers may not follow a consistent pattern. Prior studies with industrial applications have often relied on the fundamental frequency from the Fourier transform or power spectral density of the inertial signals as the indicator of the task repetitions (Radwin and Lin, 1993; Douphrate et al., 2012; Álvarez et al., 2016). An example of one of the few studies to apply a robust approach is Peppoloni et al. (2016) where they incorporated the signals from multiple IMUs and EMGs in a state transition machine to detect each cycle of cashiers' item scanning task. Park and Kim (2017) applied CWT on the data from depth camera to capture task cycles by capitalizing on the time-frequency domain features. In our estimation, the time-frequency domain features should be preferred in industrial applications where the pace of the tasks/sub-tasks can change over time due to multiple

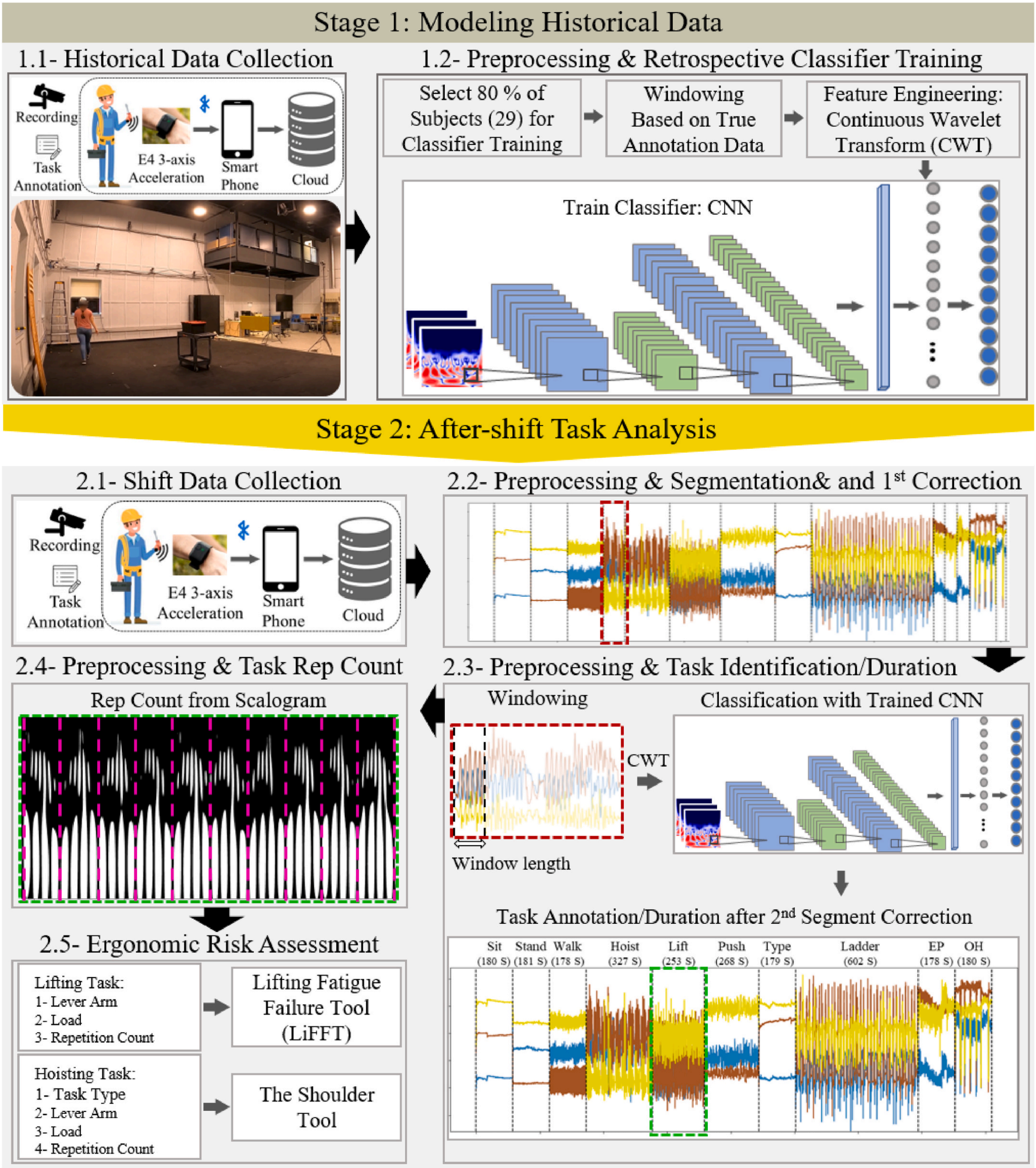
Stage 1: Modeling Historical Data
1.1- Historical Data Collection
Recording
Task Annotation
E4 3-axis Acceleration
Smart Phone
Cloud
1.2- Preprocessing & Retrospective Classifier Training
Select 80 % of Subjects (29) for Classifier Training
Windowing Based on True Annotation Data
Feature Engineering: Continuous Wavelet Transform (CWT)
Train Classifier: CNN
Stage 2: After-shift Task Analysis
2.1- Shift Data Collection
Recording
Task Annotation
E4 3-axis Acceleration
Smart Phone
Cloud
2.2- Preprocessing & Segmentation& and 1st Correction
2.3- Preprocessing & Task Identification/Duration
Windowing
Classification with Trained CNN
CWT
Window length
Task Annotation/Duration after 2nd Segment Correction
Sit (180 S)
Stand (181 S)
Walk (178 S)
Hoist (327 S)
Lift (253 S)
Push (268 S)
Type (179 S)
Ladder (602 S)
EP (178 S)
OH (180 S)
2.4- Preprocessing & Task Rep Count
Rep Count from Scalogram
2.5- Ergonomic Risk Assessment
Lifting Task:
1- Lever Arm
2- Load
3- Repetition Count
Lifting Fatigue Failure Tool (LiFFT)
Hoisting Task:
1- Task Type
2- Lever Arm
3- Load
4- Repetition Count
The Shoulder Tool

Fig. 2. An overview of our methodological framework. After training/tuning of the models in stage 1, the models are used for automated ergonomic risk assessment of worker's shift data in stage 2. (CWT: continuous wavelet transform, CNN: convolutional neural networks).

factors such as fatigue (Lamooki et al., 2021; Baghdadi et al., 2018). Simple frequency metrics such as fundamental frequency cannot account for that.

2.3. Technical gaps and contributions

The proposed framework addresses the following four gaps in the literature:

- The majority of the works in the literature investigate task identification, time series segmentation, and task duration/repetition count separately. In our framework these three pieces are implemented collectively and in accordance with each other in order to construct an end-to-end framework.
- CNN algorithms have been shown to improve the performance of classification problems since the feature extraction is built into the network architecture. The performance of task identification can also benefit from different feature engineering techniques (e.g., time-frequency features). As such, we used CWT to capitalize on time-frequency features within the CNN architecture.
- Many of the works on task repetition counting lack robustness due to failing to account for changing pace, being designed for a specific task, equating certain posture changes with task cycles, assuming global patterns of task performance, or requiring construction of personalized task profile repositories. Our task repetition counting approach is advantageous by: (a) accounting for varying pace by using the time-frequency features, and (b) allowing for application on different tasks with minimal changes (after the segmentation and task identification steps are performed).
- For the algorithms that utilize sliding windows, the performance of task identification and duration estimation depend on the choice of windows. To improve the selection of windows, we incorporated a segmentation step to detect the task transitions in the acceleration signals.

3. Methods

Fig. 2 depicts the detailed technical steps needed to translate the conceptual framework to ergonomic practice. The methods are divided into two stages. In stage 1, “historical” data are collected for one-time training of a CNN classifier for task identification as well as hyper-parameter tuning of the segmentation and task repetition counting models. The tasks were performed by a group of participants while data were captured using an accelerometer on the dominant wrist. This offline model training eliminates the need for retraining the model for new/unseen subjects. Hence, we use the term “historical” throughout the paper to differentiate between the training data and the data used in stage 2 for “after-shift task analysis”. In stage 2, we perform the ergonomic risk assessments for a given operator based on their exposures over an entire work period. Stage 2 utilizes the trained/tuned models for segmentation, task identification, and estimation of task duration/repetition counts as inputs to ergonomic risk assessment tools.

3.1. Data collection

We recruited 37 participants with the demographics summarized in Table 1 to perform 10 tasks that simulated those commonly performed by electrical line workers (Vergara et al., 2020). The tasks are described in detail in Table 2, with the order of performance as followed by the rows. The study was approved by the University at Buffalo Institutional Review Board and all participants provided signed informed consent prior to participating. The participants were equipped with an Empatica E4 wristband (Empatica, Boston, United States) on their dominant wrist to collect the 3-axis acceleration data at a sampling rate of 32 Hz (Fig. 3). Using the E4 real-time app on a smartphone carried in the participant’s pocket, the data were collected via Bluetooth connection and automatically uploaded to the cloud at the end of each session. An annotator recorded the start and end times of each task during the session to serve as the ground-truth for our segmentation and task identification models. The unlabeled data were excluded from our analysis and the annotated data were preserved and concatenated to be used in the proposed framework.

Table 1
Summary of anthropometric data for the participants.

Gender	Count	Age	Body Mass (kg)	Height (m)
Male	23	27.6 ± 4.2	80.1 ± 15.3	1.80 ± 0.08
Female	14	23.4 ± 5.1	63.3 ± 14.5	1.63 ± 0.07

Table 2
Detailed description of the tasks performed in the order followed by the rows.

Task name	Task description
Sitting	Sitting on a chair with the hands resting on the chair arm for 3 min
Standing	Standing for 3 min
Walking	Walking on set path for 3 min
Hoisting	Hoisting a 2.5 kg bucket attached to a rope up to a height of 3 m followed by hoisting down to the ground for 10 times
Lifting	Reaching for a 10 kg box on a cart positioned in front of the subject, lifting down the box to the ground, unbending to straight pose, bending and reaching for the box on the ground, and lifting the box back on the cart
Pushing	Pushing a cart (loaded with a 10 kg box) on a 10 m path, turning at the end of the path and continue pushing back to the start point for 10 times
Typing	Typing on a keyboard for 3 min while sitting on a chair
Ladder	Ascending and descending a ladder for 20 times
Overhead	Inserting screws to holes at overhead level (adjusted based on subject’s height) using a screw driver
Electrical panel	Standing in front of a simulated electric panel and performing the designed task using the dominant hand

3.2. Modeling historical data

We randomly selected 29 (about 80%) of the 37 participants and used their acceleration data for the training/parameter tuning in stage 1. In the following subsections, we describe the data processing specific to each step.

3.2.1. Retrospective classifier training

We smoothed the wrist acceleration signals using a 4th order, zero-lag, low-pass Butterworth filter with a cut-off frequency of 2 Hz. Next, we extracted 10-s non-overlapping sliding windows from the acceleration signals. Across the 29 training subjects there were 6184 observations (windows). The signals in each window were transformed to the

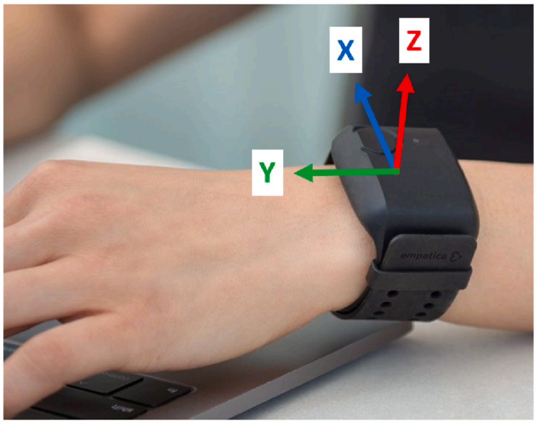

Fig. 3. The placement of the Empatica E4 wristband and the accelerometer axis orientation.

time-frequency domain using Continuous Wavelet Transform (CWT) with scales from 1 to 199 and Morlet mother wavelet. Three wavelet coefficient matrices were constructed, each associated with one sensor channel. CWT provided 2D representation of the signals used in the CNN architecture for task classification/identification. The three wavelet coefficient matrices were stacked on the third dimension to create a 3D array as the input to the CNN with 199 × 320 × 3 input shape. The three dimensions correspond to the wavelet scales (frequencies), window length, and the three accelerometer channels, respectively. The CWT feature extraction technique was selected to further construct signal scalograms used for task repetition counting in stage 2.4.

In the CNN architecture, after the input layer there are 2 convolutional layers with 32 and 64 filters. Each convolutional layer had a kernel size of 5 × 5 with ReLU activation function and is followed by a max-pooling layer. At the end there is a fully connected layer with 100 neurons followed by a softmax layer to output the probability of each of the 10 classes (tasks). This is a standard and relatively simple architecture that is similar to the widely used AlexNet (Krizhevsky et al., 2012). We used categorical cross-entropy loss function with Adam optimizer (learning rate of 0.001) and batch size of 20 for training. The CNN architecture is shown in Fig. 4. The performance of the classifier is reported in terms of average training accuracy in Section 4.1 and task- and subject-specific precision, recall, and $f1 - $ score for the validation (holdout) set in Section 4.2.2.

3.3. *Modeling shift data*

The overarching goal of this stage is to quantify the risk factors used in the ergonomic risk assessment tools. To achieve this goal, we identify each task's start and end (segmentation), annotate the task within a given segment (task identification), and record the task duration/number of repetitions. Prior to each of the above steps the acceleration signals were smoothed using a 4th order, zero-lag, low-pass Butterworth filter with a cut-off frequency of 2 Hz. We used the data from the remaining 8 subjects (about 20%) to evaluate the performance of the models used in our "after-shift" data analysis (stage 2).

3.3.1. *Segmentation*

We used the Greedy Gaussian Segmentation (GGS), a multivariate change point detection algorithm (Hallac et al., 2019) to detect the task transitions. Multivariate segmentation was selected over univariate approaches (Lavielle, 2005), because the change points may appear in different sensor channels for different task transitions. For instance, if walking is immediately followed by pushing, the $x -$ acceleration signal (blue signal in stage 2.3 of Fig. 2) may not be useable to detect the transition using a univariate segmentation approach.

The GGS algorithm detects the segment boundaries based on the assumption that the data in each segment are from the same multivariate Gaussian distribution with a mean and covariance that are independent of other segments. The algorithm achieves a locally optimal solution that scales linearly with the time series length. The algorithms that solve for the global optimum use dynamic programming (Kehagias et al., 2006; Fragkou et al., 2004) with complexities that grow with square of the time series length, thus being computationally inefficient for the worker's shift-long data. In GGS, a parametrized distribution is defined for the segmented time series (with K segments), called Segmented Gaussian Model (SGM). The maximum log-likelihood of the SGM is estimated while regularizing the covariance with λ as the regularization parameter at a fixed number of segments. The model has two hyperparameters, λ and K that are tuned based on the historical data.

To improve the GGS segments in our application, two corrections were applied on the detected segments in stages 2.2 and 2.3 (Fig. 2). By assigning the exact number of segments to K (for the 10 activities) on the historical data, we observed that while in some cases all transitions were detected correctly, at times, multiple segment boundaries were detected near one task transition, thus missing some others. By increasing K, all transitions were detected; however, this adds extra segment boundaries to the results. It should be noted that the GGS algorithm iteratively adds breakpoints in the time series (from 1 to K). To only preserve one segment boundary in close proximities (30 s), we searched the GGS outputs in the iterations from 1 to K and retained the one that appeared in the earliest iteration (first correction).

Segmentation performance was evaluated using the covering metric in (van den Burg and Williams, 2003). It is calculated as

$$C(S^{'}, S) = \frac{1}{T}\sum_{A\in S}|A| \, . \max_{A^{'}\in S^{'}} J(A, A^{'}),$$

where T is the length of the time series, and S and S' are the set of true and detected segments, respectively. For each of the true segments, their Jaccard indices (J, intersection over union) are calculated against all detected segments. Next, the covering metric was obtained as the weighted average of maximum Jaccards of the true segments. The covering metric was calculated after initial segmentation and each correction.

3.3.2. *Task identification/duration*

We created 10-s non-overlapping sliding windows using the detected segment boundaries. Next, we extracted the wavelet coefficients by CWT on each window and classified the windows using the trained CNN. We assign one task to each segment by majority voting among the windows within the segment. As the second segmentation correction, the adjacent segments with the same labels were merged together. These were the

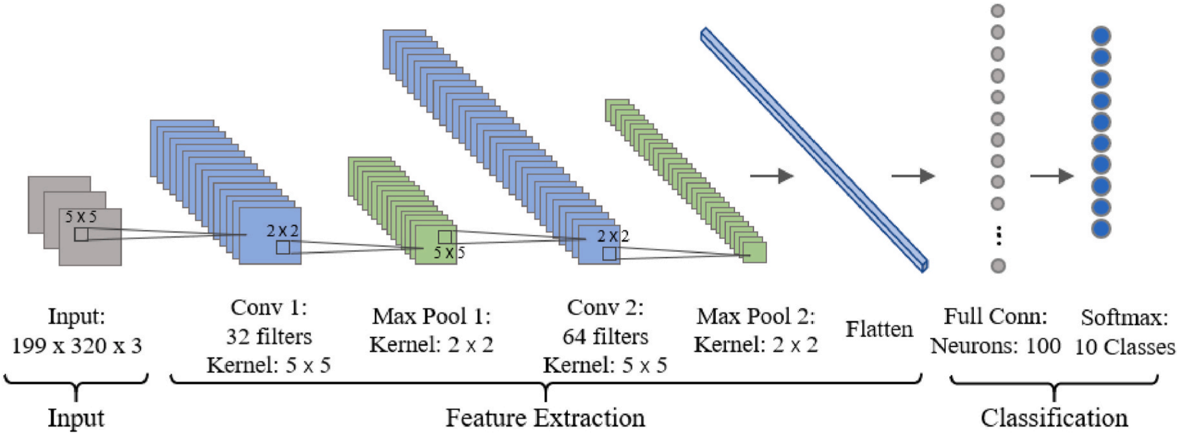

Fig. 4. The architecture of our CNN with an input layer, two convolutional layers (each followed by a max-pooling layer), followed by a fully connected layer with 100 neurons and the softmax layer for multiclass predictions.

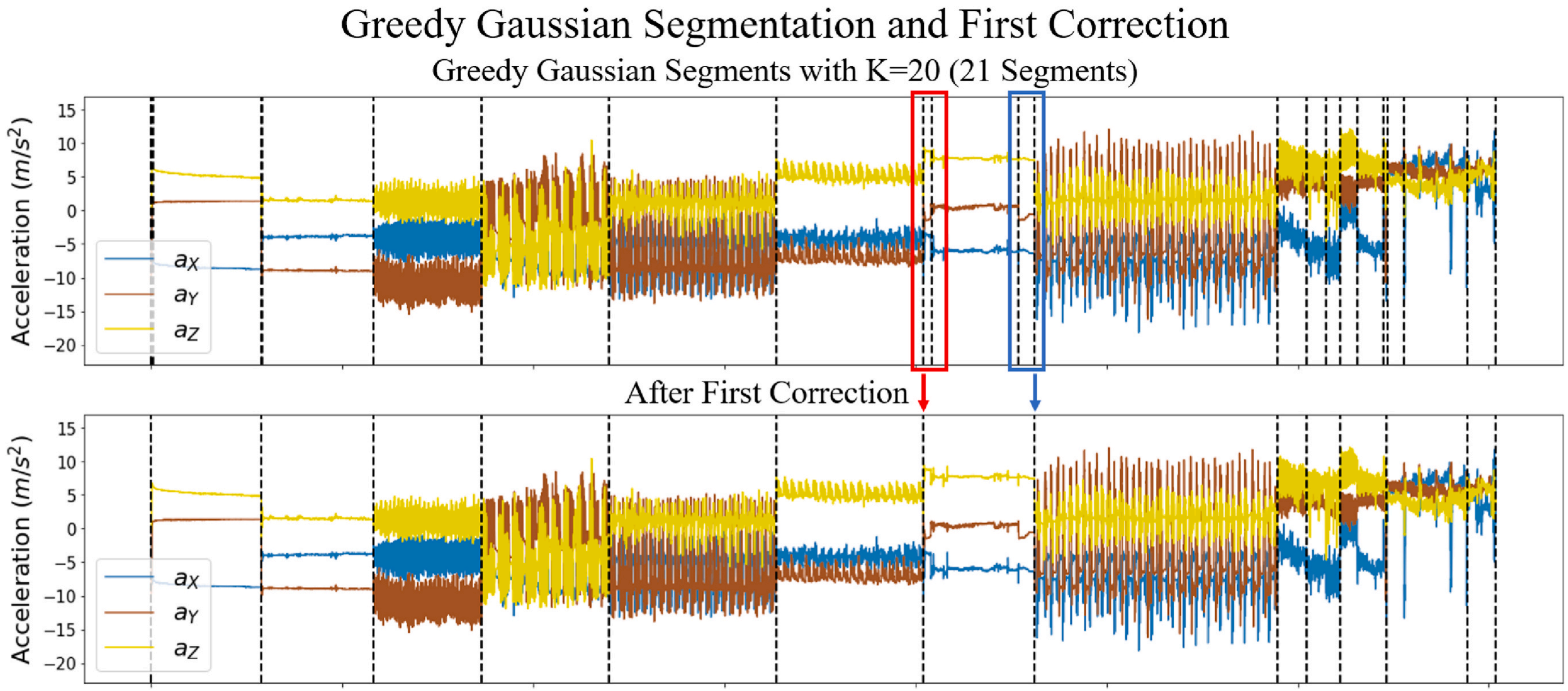

Fig. 5. Initial segmentation and first correction are presented for representative test subject 8. The top panel shows the GGS segmentation results set to find 20 segment boundaries (excluding start and end). First correction (bottom panel) preserves the best segment boundary from a group of segments in close proximity of each other (2 examples are highlighted in red and blue boxes). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

final segments and used to calculate the task duration/number of repetitions. For evaluation of the task identification performance, we investigated the classifier's predictive performance on each window using the precision, recall, and $f1-$ score. Furthermore, we use the $overall\, accuracy = \frac{True\ Positives + True\ Negatives}{Number\ of\ observations}$ for an overall assessment.

3.3.3. Task repetition counting

For counting the number of repetitions for a given task, the filtered acceleration signals were converted to scalograms. A major axis was then selected for the repetition counting model, depending on the kinematic characteristics of the task to preserve the most relevant information. To detect individual task cycles, we converted the scalograms to binary images by a threshold based on the distribution of the absolute wavelet coefficients. With this approach, the task cycles appear as separate contours (isolated regions) within certain wavelet scale intervals, specific to the task. The contours can correspond to different components of one task cycle (i.e., sub-tasks). For example reaching for the rope and pulling up the rope are the two components in one hoisting repetition. Depending on the definition of one task repetition, a certain number of contours will represent one task repetition. This approach allows for repetition counting in non-stationary signals where task pace may change.

The major axis and the model hyperparameters (mother wavelet and threshold), are selected/tuned for each task. We selected the major axis by evaluating the performance of each accelerometer channel on the historical data. For this study, we applied the method to count the number of repetitions during the lifting and hoisting tasks. These tasks were selected since they loaded the low back and shoulder, respectively, two joints of common interest for ergonomics risk assessment (Nordander et al., 2016; Antwi-Afari et al., 2017). The $y-$ acceleration signal performed better than x and z for both tasks and was used as the major axis. With minor modifications/tuning, the model can be applied to other activities, such as walking for counting the number of steps. To evaluate the performance of this approach, we calculated the error rate for the predicted vs true number of task repetitions.

3.3.4. Ergonomic risk assessment

The output of stages 2.2–4 include task label, task duration, and/or number of repetitions. Task label and duration can be directly used for ergonomic assessment of the shift data and the number of repetitions is required in ergonomic tools for risk assessment. We used LiFFT and the Shoulder Tool for ergonomic risk assessment in this application to investigate the cumulative exposure of the engaged tissues using the fatigue failure theory (Gallagher et al., 2017; Bani Hani et al., 2021). LiFFT evaluates the risk of back loading during the lifting task where the low back tissues are exposed to cumulative damage. For risk assessment during the hoisting task, we used the Shoulder Tool which estimates the cumulative damage to the shoulder tissues. Both tools require the number of task repetitions, lever arm, and load as inputs. For the lever arm in LiFFT, we used the arm length, estimated from the participant's height (Bass, 1987), since the participant reached to full arm length as the maximum horizontal distance when placing the box on the cart during the lift. For the Shoulder Tool we used half of the arm length as the lever arm, since based on our observations from the recorded videos, the subjects' shoulder distance from the rope was about half of their arm length. Also, a 10 kg (22.05 lb) box and a 2.5 kg (5.51 lb) bucket were used for the lifting and hoisting tasks, respectively.

4. Results

4.1. Modeling historical data

Across all subjects and tasks, the model achieved an average overall training accuracy of 98%. Due to the excellent predictive performance and for the sake of conciseness, we do not further detail how the models performed on our training set. The results of 8 test subjects are reported in the following sections.

4.2. After-shift task analysis

4.2.1. Segmentation

We evaluated the segments from the initial GGS application and after the first correction in terms of covering metric in the first two rows of Table 3. Note that a covering metric of 1 indicates perfect segmentation. Prior to the correction, the covering metric ranged from 0.739 (subject 3) to 0.885 (subject 1), with an average value of 0.840. After the first correction, the covering metric improved, ranging from 0.751 (subject 3) to 0.969 (subject 1), with an average of 0.865.

To illustrate our approach for GGS application and the first correction, we depict the initial and corrected segment boundaries for a representative subject in Fig. 5. The red and blue boxes show two examples of multiple segment boundaries detected near one task transition. As shown in the figure, the first correction preserves the segment boundary that captures the true task transition.

4.2.2. Task identification and duration estimation

In Table 4, we present the precision, recall, and $f1-$ score for task-specific evaluation, and the subject-specific overall accuracy to capture the performance of the CNN classifier across all tasks. There are three main observations. First, the overall accuracy for 7 out of the 8 subjects is excellent ($\geq$ 92%, with an average of 96%). Second, the task identification results are significantly lower for subject 3 who was left handed. Their poor task identification result will propagate hereafter and adversely affect the second segmentation correction and also diminishes the task duration/repetition counting performance. Third, if we were to exclude subject 3 and focus on the $f1$ results, the $f1-$ scores for our 10 tasks are 98% (electrical panel), 98% (hoisting), 96% (ladder), 97% (lifting), 100% (overhead), 99% (pushing), 90% (sitting), 98% (standing), 86% (typing) and 100% (walking). For these seven subjects, the CNN classifier can accurately detect all tasks, with perfect annotation of the overhead and walking tasks. The performance of the CNN is worst for both the typing and sitting tasks due to the recall for the typing task being 80% and the precision for the sitting task being 86%.

To provide additional insights on how the CNN classifier performed, we present the confusion matrices for test subjects 1 and 3 in Fig. 6, where the rows and columns denote the true and predicted labels, respectively. For subject 1, only two of the tasks had windows that were misclassified, where 3 (out of 24) sliding windows of ladder were predicted as lifting and 1 (out of 18) windows for sitting was classified as typing. On the other hand, for subject 3, there were a larger number of instances where our sliding windows were misclassified. If we were to focus on the diagonal, the ladder, pushing and walking tasks had the worst predictive performance.

The third row in Table 3 presents the covering metric after the second segmentation correction. The second correction, which is based on the task identification results (by assigning one task/label to each segment and merging adjacent segments with the same label), further enhanced the segmentation. The final overall average covering metric improved from 0.865 to 0.979 (i.e., 13% improvement). Moreover, the covering metric improved for all subjects, with the largest improvement for subject 3 (from 0.751 to 0.892, 19% improvement).

Based on the labeled segments, the tasks' durations were computed from the start and end segment boundaries. The labeled segments, with their corresponding durations, are presented in Fig. 7 for the two representative test subjects. For the right-handed subjects, the mean duration errors remained below 1% with an average of 0.6%, while the error for the left-handed subject was at 53.8%. The incorrectly labeled segments for this subject significantly contributed to the duration error, with the largest error associated with the "overhead" task (125%) as the entire "pushing" segment was misclassified as "overhead". The misclassified segments also caused large errors for estimated duration of the

Table 3
Segmentation performance using covering metric.

Test subjects		1	2	3	4	5	6	7	8	Average
Covering Metric	Initial GGS segments	0.885	0.877	0.739	0.817	0.869	0.864	0.793	0.878	0.840
	Segments after first correction	0.969	0.889	0.751	0.833	0.886	0.864	0.807	0.923	0.865
	Segments after second correction	0.998	0.997	0.892	0.992	0.993	0.987	0.986	0.99	0.979

Table 4
Classification performance for each test subject during 10 tasks (EP: electrical panel, H: hoisting, Ld: ladder, Lf: lifting; OH: overhead; P: pushing; St: sitting; Sd: standing; Tp: typing; W: walking), The performance is reported as p: precision, r: recall, and f1: $f1 -$ score.

Test Subject	Task-specific Performance										Subject Overall Accuracy
		EP	H	Ld	Lf	OH	P	St	Sd	Tp	W	
1	p	1.00	1.00	1.00	0.87	1.00	1.00	1.00	1.00	0.95	1.00	0.98
	r	1.00	1.00	0.88	1.00	1.00	1.00	0.94	1.00	1.00	1.00	
	f1	1.00	1.00	0.93	0.93	1.00	1.00	0.97	1.00	0.97	1.00	
2	p	0.90	0.97	1.00	0.93	1.00	0.96	0.67	0.82	0.83	1.00	0.92
	r	1.00	0.97	0.85	1.00	1.00	1.00	0.89	1.00	0.56	1.00	
	f1	0.95	0.97	0.92	0.96	1.00	0.98	0.76	0.90	0.67	1.00	
3	p	0.25	0.91	0.74	0.67	0.19	0.00	1.00	1.00	1.00	1.00	0.60
	r	0.44	0.91	0.33	0.87	0.50	0.00	0.94	1.00	1.00	0.28	
	f1	0.32	0.91	0.46	0.75	0.28	0.00	0.97	1.00	1.00	0.43	
4	p	0.94	1.00	0.93	1.00	1.00	0.96	1.00	1.00	0.94	1.00	0.97
	r	0.94	1.00	1.00	0.92	1.00	1.00	0.94	0.94	0.94	1.00	
	f1	0.94	1.00	0.96	0.96	1.00	0.98	0.97	0.97	0.94	1.00	
5	p	1.00	1.00	0.98	1.00	1.00	1.00	1.00	0.95	1.00	1.00	0.99
	r	1.00	1.00	0.98	0.96	1.00	1.00	1.00	1.00	1.00	1.00	
	f1	1.00	1.00	0.98	0.98	1.00	1.00	1.00	0.97	1.00	1.00	
6	p	1.00	1.00	0.98	1.00	1.00	1.00	0.78	1.00	1.00	1.00	0.97
	r	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	0.72	1.00	
	f1	1.00	1.00	0.99	0.98	1.00	1.00	0.88	1.00	0.84	1.00	
7	p	1.00	0.95	0.93	1.00	1.00	1.00	0.89	1.00	1.00	1.00	0.97
	r	1.00	0.88	1.00	1.00	1.00	1.00	0.94	1.00	0.89	1.00	
	f1	1.00	0.91	0.96	1.00	1.00	1.00	0.92	1.00	0.94	1.00	
8	p	1.00	1.00	0.95	1.00	1.00	1.00	0.67	1.00	1.00	1.00	0.95
	r	0.94	1.00	1.00	0.96	1.00	1.00	1.00	1.00	0.50	1.00	
	f1	0.97	1.00	0.98	0.98	1.00	1.00	0.80	1.00	0.67	1.00	

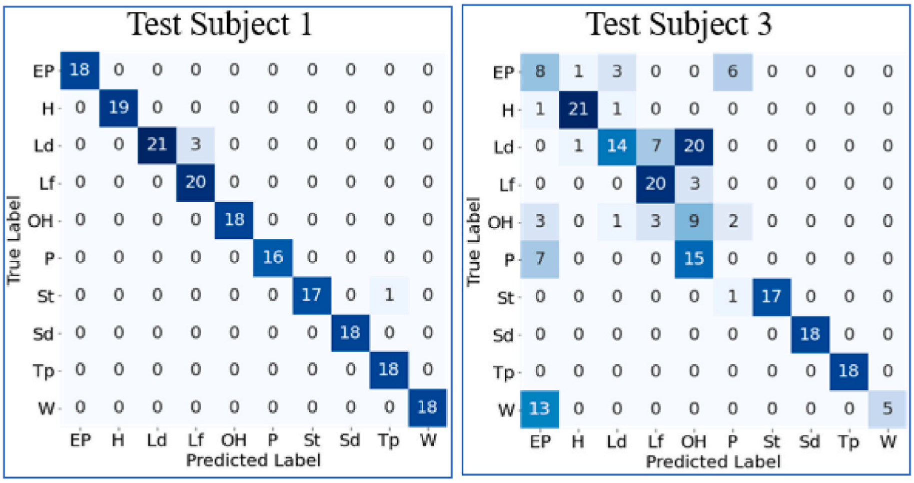

Fig. 6. Confusion matrices for subjects 1 and 3. (EP: electrical panel, H: hoisting, Ld: ladder, Lf: lifting; OH: overhead; P: pushing; St: sitting; Sd: standing; Tp: typing; W: walking).

electrical panel, ladder, pushing, and walking tasks (above 45% error). The mean duration error across the 8 test subjects and all activities was 7.3%.

4.2.3. *Task repetition counting*

4.2.3.1. Lifting. We depict how the scalogram of the $y -$ acceleration was used for counting the repetitions for the lifting task for a representative subject (test subject 1) in Fig. 8. The figure highlights how the 20 repetitions of the lifting task were captured (0% error in this case). The figure also shows how both the acceleration signal and scalogram captured the repeated hand motion during the lifting task. One lifting repetition denotes a sequence of sub-tasks as described in Table 2.

4.2.3.2. Hoisting. Fig. 9 presents the scalogram of the $y -$ acceleration

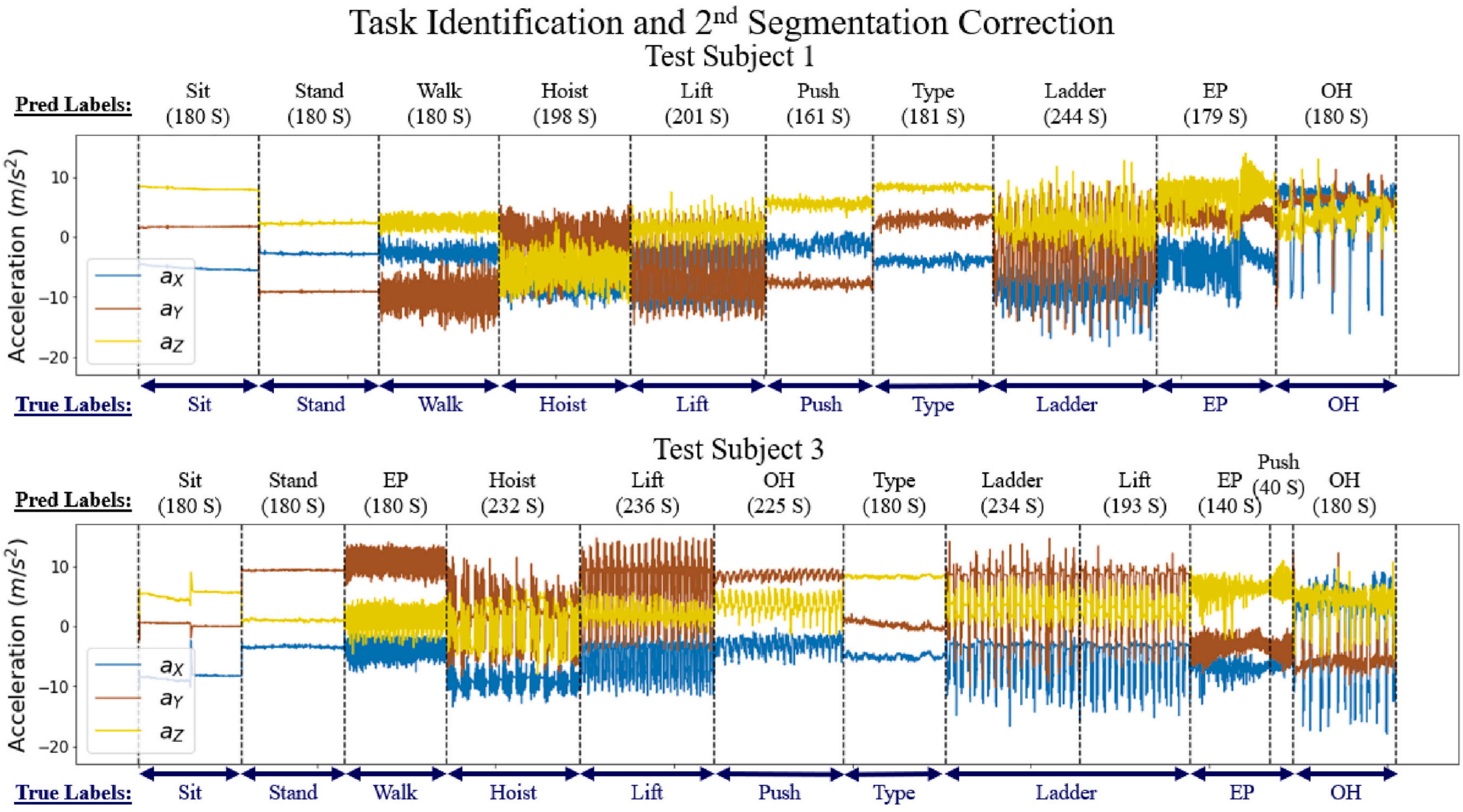

Fig. 7. Output of task identification, segment task annotation, and second segmentation correction. (EP: electric panel, OH: overhead). The vertical dashed lines present the detected segment boundaries and the arrows above the true labels demonstrate the true task start and end points.

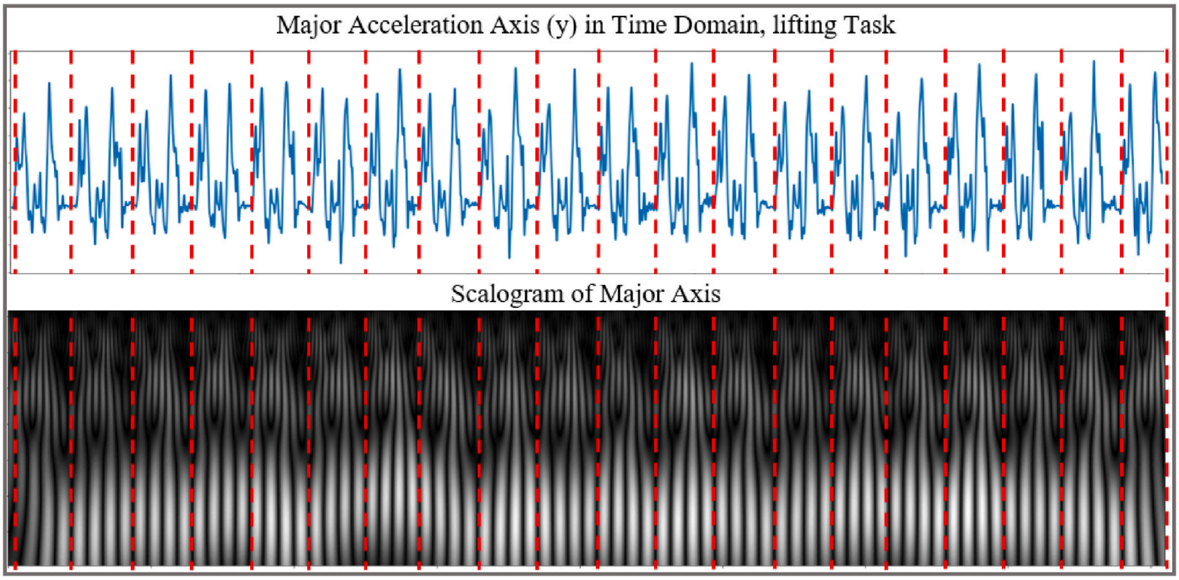

Fig. 8. Counting the number of repetitions of the lifting task using the scalogram of the major signal ($y-$ acceleration). The top panel shows the y signals in time and the bottom panel visualizes its scalogram in gray scale. Dashed vertical red lines separate the task repetitions. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

during the hoisting task for the test subject 1. One repetition of the hoisting task is defined as reaching for and pulling up the rope once (or the opposite during hoisting down). The hoisting task provides an example where the repetitions occur at a relatively higher frequencies as compared to the lifting. Here, we present both the signal and scalogram for the entire hoisting task and a zoomed-in view showing one cycle of hoisting up to the pre-specified height followed by hoisting down to the ground. From the figure, each two successive white contours in the scalogram captured one repetition. The scalogram captured 83 repetitions, from 90 repetitions performed by the participant (error rate: 8.4%).

In Table 5, we summarize the counting errors across the test subjects for both tasks. The true repetition counts for each of the lifting and hoisting tasks were obtained from the recorded videos. For lifting, our algorithm counted at most one extra/fewer repetition for all participants, except for subject 3 (with a perfect count for five of the subjects). The count was erroneous for subject 3 due to the propagation of errors from the task identification stage, with an overestimated count of 31 total lifts. This is an expected result since the poor CNN performance misidentified the majority of the ladder windows and incorrectly labeled a portion of ladder as lift (Fig. 7). Hence, this “extra” segment resulted in counting 11 additional lifts. For hoisting, the repetition counting error was relatively low for all eight subjects ($\leq 9.1\%$). There was no propagation of errors for hoisting since the CNN accurately classified the hoisting task ($f1$ of 0.91) for subject 3 (Table 4).

4.2.4. *Ergonomic risk assessment*

The results from the previous sections were then used as inputs to the fatigue failure-based ergonomic risk assessment tools. The results are presented in Table 6 using the true and predicted number of repetitions for the 8 test subjects. The predicted injury probabilities (reported in percent per the tools) varied from 11.3 to 16.0% and $7.2-8.4\%$ for the lifting and hoisting tasks, respectively.

5. Discussion and conclusions

5.1. *Summary of main contributions*

For assessment of physical exposure in industry, a myriad of tools are widely used in practice (e.g., REBA, RULA, OWAS, LiFFT, the Shoulder Tool (Hignett and McAtamney, 2000; McAtamney and Corlett, 1993; Karhu et al., 1977; Gallagher et al., 2017; Gallagher et al., 2018; Bani Hani et al., 2021)). However, such tools typically require monitoring of multiple risk factors by professionals (direct observation), thus limiting their feasibility for large-scale assessment (Takala et al., 2010; van der Beek and Frings-Dresen, 1998). In this study, we have shown the utility of an end-to-end framework and methodology for the automated monitoring of three important ergonomic risk factors: (a) task type; (b) task duration; and (c) task repetition count. We integrated this information to quantify the ergonomic risk associated with the back (due to lifting) and shoulder (due to hoisting) joints using LiFFT and the Shoulder Tool, respectively. It is important to note that the evaluation of our framework steps was based on an inter-subject scenario (i.e., population level) to avoid re-training for unseen subjects.

In Section 4, we have shown that our modified GGS implementation results in excellent segmentation performance, which was improved with our first correction (30-s window) and the final correction (segment merging based on the CNN approach). So one may wonder why segmentation is needed when a multi-class classification task identification model may be used. In our estimation, there are two main practical scenarios where the use of both segmentation and classification may be superior to a single approach. First, the use of statistical multiple change-point detection methods for segmentation would commonly result in segment lengths that are larger than the sliding window lengths used in the classification models. By ensembling both results, we can reduce the false alarm rate from the classification models (i.e., improve the overall specificity for those models). Second, segmentation methods can improve the estimation of task duration in applications where the window size used for classification has to be relatively large and/or mixed task scenarios (where a task is repeated but not performed sequentially). While we did not consider the mixed task scenario in this article, our results show how the overall task identification and duration results improved by combining the segmentation and CNN outputs (see Fig. 7).

Based on our 8 unseen (holdout/testing) subjects, we achieved an overall average accuracy of 92.0% for task recognition (tested on all 10 tasks), 7.3% error for task duration (tested on 8 tasks), and 7.1% error for task repetition counting (tested on 2 tasks). The results are consistent with Kim and Nussbaum (2014) (6 tasks, 17 IMUs, with an average recall of 80% and underestimating task duration by ~14% using a multilayer

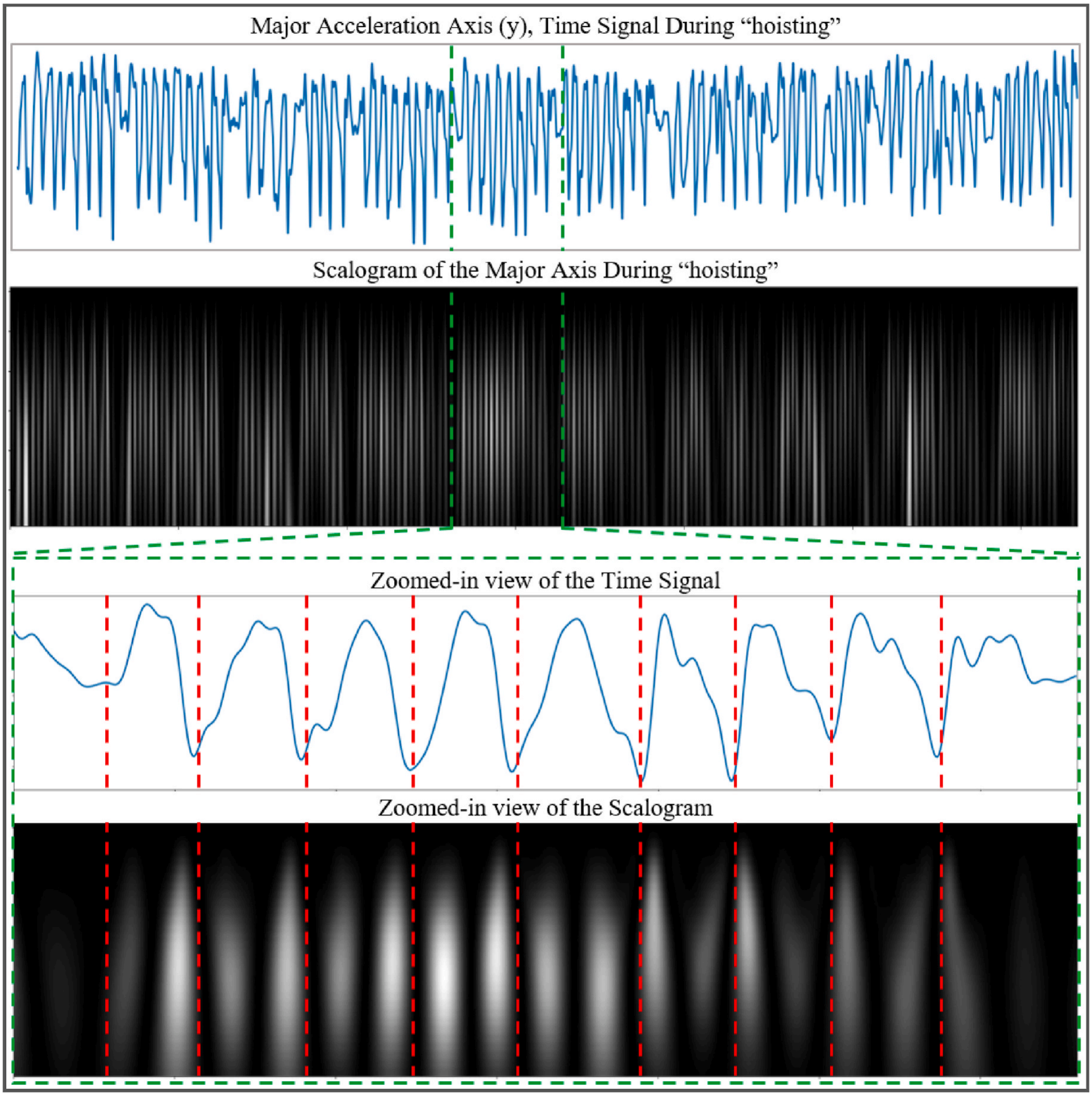

Fig. 9. Counting the number of repetitions for the hoisting task using the scalogram of the major axis ($y -$ acceleration). The time signal is shown as blue curve. The top panel includes the entire hoisting performance for a representative subject, and the bottom panel shows a zoomed-in view and separates individual repetitions of the task by the dashed vertical lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 5
Repetition counting for the "lifting" and "hoisting" tasks for the 8 test subjects.

Test Subject	Lifting			Hoisting		
	True Count	Predicted Count	Error Rate (%)	True Count	Predicted Count	Error Rate (%)
1	20	20	0.0	83	90	8.4
2	20	19	5.0	147	142	3.4
3	20	31	55.0	88	96	9.1
4	20	20	0.0	92	90	2.2
5	20	20	0.0	101	94	6.9
6	20	20	0.0	89	94	5.6
7	20	20	0.0	80	86	7.5
8	20	21	5.0	132	124	6.1

feedforward neural network and two other classification algorithms) and Hosseinian et al. (2019) (15 tasks, single sensor, with an accuracy of 93%–98% obtained using support vector machine and random forest models), which are the two studies with somewhat similar tasks, summarized in the review of (Lim and D'Souza, 2020). Note that the classification results cannot be compared directly since the performed tasks are not identical, vary in complexity and different types of sensors were used.

While our results are consistent with previously reported task identification studies, our work builds upon those studies in three ways. First, we examined several complex tasks, where each of these tasks is comprised of multiple sub-tasks. The complexity of the tasks required us to increase our sliding window size to 10 s to "guarantee" capturing an entire cycle of the most complex tasks. In previous studies, the window sizes typically varied from 2 − 6 s. We have shown that these complex tasks can be classified accurately. Second, with the exception of Kim and Nussbaum (2014), the task identification methods did not consider task duration and none considered repetitions beyond threshold-based counts of postures. Here, we have shown that the use of CNN with features extracted from the Continuous Wavelet Transform (CWT) allowed for accurate achievement of task identification, duration, and repetition

Table 6
The outputs of the LiFFT and The Shoulder tools for the lifting and hoisting tasks, respectively. 22.05 lb and 5.51 lb loads were used for the lifting and hoisting tasks, respectively.

Test Subject	LiFFT					The Shoulder Tool				
	Lever Arm (in)	Repetitions		Probability of High Risk Job (%)		Lever Arm (in)	Repetitions		Probability of Shoulder Outcome (%)	
		True	Pred	True	Pred		True	Pred	True	Pred
1	24.82	20	20	13.2	13.2	12.41	83	90	7.0	7.2
2	24.31	20	19	13.2	13.2	12.16	147	142	8.5	8.4
3	25.54	20	31	13.2	16.0	12.77	88	96	7.4	7.6
4	25.42	20	20	13.2	13.2	12.71	92	90	7.5	7.4
5	25.30	20	20	13.2	13.2	12.65	101	94	7.7	7.5
6	24.52	20	20	13.2	13.2	12.26	89	94	7.1	7.2
7	27.23	20	20	14.7	14.7	13.61	80	86	7.6	7.8
8	22.17	20	21	11.3	11.3	11.08	132	124	7.4	7.2

counts. Note that the application of simple peak-detection approaches for repetition counting are highly sensitive to hyperparameters that impedes their global tuning across different scenarios. Moreover, they are less effective for complex tasks with varying (multi-)peak patterns in the time domain and result in a degraded performance. Third, the scalograms of the signal from the CWT allow for reliable task repetition counting even if the pace of task performance changes. This can be important in ergonomic applications, where fatigue is of primary interest (Baghdadi et al., 2021).

Using the outputs from task identification, duration estimation and repetition counting, we have computed the probability of high risk job/task to induce: (a) back damage during the "lifting" task, and (b) shoulder damage during the "hoisting" tasks, using LiFFT and the Shoulder Tool, respectively. The estimated risk probabilities from outputs resulted in average errors of 2.6% for the back and 1.9% error for the shoulder across the 8 test subjects. These error rates are insignificant in practice since they are likely within the error margins of the tools. Note that, due to using fatigue-failure tools to compute ergonomic risk, we only examined task repetition counting for the lifting and hoisting tasks as two representative examples. Our task repetition counting can be applied and tuned to other non-stationary tasks (i.e., our other tasks excluding sitting and standing). For example, we could have easily counted the number of steps in walking.

Our inter-subject design of the framework eliminates re-training for unseen subjects; however, the inter-subject variabilities introduce further challenges to the framework. The framework performance was reported for two subjects with the best and worst results throughout Section 4. Task identification achieved an average accuracy of 96% across seven test subjects and dropped to 60% for one who was left-handed. Among the total 37 subjects in our study, two were left-handed with one in the training (historical) and one in the test (after-shift) data. Despite robustness of the time-frequency features to sensor orientation (Ravi et al., 2016), the sensor placement on the left hand resulted in poor task identification in our study. Since such features capitalize on frequency components, tasks with periodic motion can be identified with reduced dependency on the sensor orientation by normalizing the signals. However, since static tasks were included in our analysis (e.g., sitting and standing), we avoided signal normalization to preserve the static component of the signals as the main discriminant of such tasks, thus reducing the robustness of the extracted features to the sensor orientation. In addition, the poor task identification adversely affected the second segmentation correction and further deteriorated task duration and repetition count. Nonetheless, in practical applications this can be mitigated by including adequate number of left-handed subjects in the training population.

5.2. An examination of the sensitivity of our segmentation and task classification performance

To explore the robustness/sensitivity of our framework results to beyond some specific parameters of our experimental setup and analysis, we have considered two scenarios: (a) what is the impact of a different task order on our segmentation performance? and (b) how does the training/testing set selection affect classification performance? The first question focuses on whether our segmentation performance may have benefited from the order of the tasks assigned to the participants (e.g., consecutive tasks may have differed enough to affect our segmentation performance). In the second question, we assess the sensitivity of our test results if a different set of participants were held out of the training data.

To assess question (a), we randomized data segments for the eight participants in our testing data. In our original experiment the participants performed the tasks in a set/fixed order, and hence, we shuffled the data associated with each task to simulate if they had performed the tasks in a different order. For example, in our follow-up analysis, the tasks performed by test subjects 1 and 3 were as follows: (i) *subject 1:* typing → ladder → walking → overhead → electric panel → standing → hoisting → sitting → pushing → lifting; and (ii) *subject 3:* walking → pushing → sitting → ladder → lifting → standing → overhead → hoisting → typing → electric panel.

Based on the randomized orders (see our Python/Jupyter Notebook provided in the Github repository for details), we present the corresponding covering metric results after the first and second corrections in Table 7. The average covering metrics across the 8 test subjects are 0.839, 0.956, and 0.976 after the initial segmentation, first correction, and second correction, respectively. Recall that the corresponding covering metrics for the original task orders were 0.840, 0.865, and 0.979, which are in good agreement with the results after reorganizing the task sequence.

To examine question (b), we have run the analysis using a different testing set that was a randomly selected set of 8 subjects, constrained on having no overlap with the original test set. The CNN model was retrained using the new set of training subjects and tested on the new test set. We achieved overall training and testing accuracies of 99% and 96%, respectively (which are similar to the overall 98% training accuracy and 92% testing accuracy achieved using the original training and testing sets). This provides some evidence that our approach and methodology can produce similar accuracy levels with different test subjects.

5.3. Limitations and suggested future directions

There are some limitations that can potentially reduce the generalizability of our results. First, our study has been performed in a laboratory environment. The performance of task identification can be overestimated in a controlled setting (van Hees et al., 2013). While the large number of subjects in our study and the minimal control imposed on the task performance can partly account for the expected heterogeneities in the workplace, a field study can better capture the variabilities in uncontrolled settings. Second, while the framework can be used to

Table 7
Segmentation performance after shuffling the task orders.

Test subjects		1	2	3	4	5	6	7	8	Average
Covering Metric	Initial GGS segments	0.858	0.874	0.733	0.83	0.876	0.866	0.795	0.878	0.839
	Segments after first correction	0.919	0.886	0.743	0.834	0.884	0.872	0.796	0.914	0.856
	Segments after second correction	0.997	0.993	0.884	0.993	0.996	0.995	0.967	0.979	0.976

capture task type and time-based risk factors, the load/force of the exertions cannot be automatically identified and has been input for risk assessment here based on prior knowledge of the box/bucket weights. Such information may not be readily available or may vary over the course of the work shift. Alternative approaches, such as shoe- or equipment-based sensors, can be investigated to automate acquisition of the handled weights as a future research direction. Third, the framework is designed for after-shift analysis of the worker's data in offline mode for adjustment of the work design. Near real-time assessment of ergonomic risk is of great value for providing feedback to the worker and should be pursued as an advancement of the framework. Fourth, as the performance of the framework is significantly affected by the sensor location (e.g., left vs. right wrist), transfer learning/domain adaptation can be used as a potential solution. For future research, the authors are exploring the application of domain adaptation to enable the effective use of different sensor orientation in the training and testing data. Finally, in this study the participants performed the time-based tasks for 3 min and the repetition-based tasks for ten or twenty repetitions. However, in practical applications different tasks can be performed intermittently and for fewer repetitions/shorter duration. This scenario (mixed task) is more challenging in terms of annotation and windowing and thus an important area of exploration for future research.

5.4. Concluding remarks

This study used a commercial wristband with a 3-axis accelerometer to evaluate the ergonomic risks associated with the data from the simulated tasks of the electrical line workers. We demonstrated the effectiveness of the proposed end-to-end framework to answer the five research questions posed in Section 1. The GGS segmentation algorithm was able to achieve an average 0.979 covering metric by implementation of two corrections (research question 1). These performance results held when task order was randomized for the test subjects. Further, we obtained >90% accuracy for activity recognition by application of majority voting within the detected segments on two different testing sets (research question 2). Also, the use of change point detection is shown to enhance the estimation of task duration and repetitions by accurate detection of task transition times (7.3% error in estimating task duration and 7.1% error for capturing the task repetitions; research questions 3 and 4). Finally, we demonstrated the accuracy of the present framework to systematically estimate the probability of the workplace injury using two existing ergonomic tools of LiFFT and the Shoulder tool (research question 5). Our results suggest that by using the wrist acceleration data in the proposed framework we can reliably calculate the risk factors for automated estimation of ergonomic risks.

Funding

This study was supported by the *Centers for Disease Control and Prevention* (1 R21OH011749-01), the *Pilot Projects Research Training Program of the NY/NJ Education and Research Center, National Institute for Occupational Safety and Health* (T42 OH008422), and *GE* Research. Its contents are solely the responsibility of the authors and do not represent the official views of the *Centers for Disease Control and Prevention* or *GE Research*.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The data collected for the manuscript was part of a study funded by GE Research. The experimental tasks were approved by collaborators at GE Research. They had no further involvement in the data collection and analysis or development of the manuscript.

Supplemental information

Data, Code and Analysis: The data, code and **Python** scripts used in this paper can be accessed at: https://github.com/saebragani/Automated-Ergonomic-Risk-Assessment.

References

Álvarez, D., Alvarez, J.C., González, R.C., López, A.M., 2016. Upper limb joint angle measurement in occupational health. Comput. Methods Biomech. Biomed. Eng. 19 (2), 159–170.

Antwi-Afari, M., Li, H., Edwards, D., Pärn, E., Seo, J., Wong, A., 2017. Biomechanical analysis of risk factors for work-related musculoskeletal disorders during repetitive lifting task in construction workers. Autom. ConStruct. 83, 41–47.

Baghdadi, A., Megahed, F.M., Esfahani, E.T., Cavuoto, L.A., 2018. A machine learning approach to detect changes in gait parameters following a fatiguing occupational task. Ergonomics 61 (8), 1116–1129.

Baghdadi, A., Cavuoto, L.A., Jones-Farmer, A., Rigdon, S.E., Esfahani, E.T., Megahed, F. M., 2021. Monitoring worker fatigue using wearable devices: a case study to detect changes in gait parameters. J. Qual. Technol. 53 (1), 47–71.

Bani Hani, D., Huangfu, R., Sesek, R., Schall Jr., M.C., Davis, G.A., Gallagher, S., 2021. Development and validation of a cumulative exposure shoulder risk assessment tool based on fatigue failure theory. Ergonomics 64 (1), 39–54.

Bass, W.M., 1987. Human Osteology: a Laboratory and Field Manual, vol. 2. Missouri Archaeological Society.

Burdick, R.K., Borror, C.M., Montgomery, D.C., 2003. A review of methods for measurement systems capability analysis. J. Qual. Technol. 35 (4), 342–354.

Camomilla, V., Bergamini, E., Fantozzi, S., Vannozzi, G., 2018. Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: a systematic review. Sensors 18 (3), 873.

Cavuoto, L., Megahed, F.. Understanding Fatigue: Implications for Worker Safety. Professional Safety December (2017) 16–19, URL. https://foundation.assp.org/docs/BPCav_1217z.pdf.

Chang, K.-h., Chen, M.Y., Canny, J., 2007. Tracking free-weight exercises. In: International Conference on Ubiquitous Computing. Springer, pp. 19–37.

Choi, K.S., Joo, Y.S., Kim, S.-K., 2013. Automatic exercise counter for outdoor exercise equipment. In: 2013 IEEE International Conference on Consumer Electronics (ICCE), vols. 436–437. IEEE.

Cortes, N., Onate, J., Morrison, S., 2014. Differential effects of fatigue on movement variability. Gait Posture 39 (3), 888–893.

Da Costa, B.R., Vieira, E.R., 2010. Risk factors for work-related musculoskeletal disorders: a systematic review of recent longitudinal studies. Am. J. Ind. Med. 53 (3), 285–323.

Douphrate, D.I., Fethke, N.B., Nonnenmann, M.W., Rosecrance, J.C., Reynolds, S.J., 2012. Full shift arm inclinometry among dairy parlor workers: a feasibility study in a challenging work environment. Appl. Ergon. 43 (3), 604–613.

Fragkou, P., Petridis, V., Kehagias, A., 2004. A dynamic programming algorithm for linear text segmentation. J. Intell. Inf. Syst. 23 (2), 179–197.

Gallagher, S., Heberger, J.R., 2013. Examining the interaction of force and repetition on musculoskeletal disorder risk: a systematic literature review. Hum. Factors 55 (1), 108–124.

Gallagher, S., Sesek, R.F., Schall Jr., M.C., Huangfu, R., 2017. Development and validation of an easy-to-use risk assessment tool for cumulative low back loading: the Lifting Fatigue Failure Tool (LiFFT). Appl. Ergon. 63, 142–150.

Gallagher, S., Schall Jr., M.C., Sesek, R.F., Huangfu, R., 2018. An upper extremity risk assessment tool based on material fatigue failure theory: the distal upper extremity tool (DUET). Hum. Factors 60 (8), 1146–1162.

Hajifar, S., Sun, H., Megahed, F.M., Jones-Farmer, L.A., Rashedi, E., Cavuoto, L.A., 2021a. 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.

Hajifar, S., Lamooki, S.R., Cavuoto, L.A., Megahed, F.M., Sun, H., 2021b. Investigation of heterogeneity sources for occupational task recognition via transfer learning. Sensors 21 (19), 6677.

Hallac, D., Nystrup, P., Boyd, S., 2019. Greedy Gaussian segmentation of multivariate time series. Adv. Data Anal. Classif. 13 (3), 727–751.

Hallac, D., Nystrup, P., Boyd, S.. GGS, GitHub Repository URL. https://github.com/cvxgrp/GGS.

Hignett, S., McAtamney, L., 2000. Rapid entire body assessment (REBA). Appl. Ergon. 31 (2), 201–205.

Hosseinian, S.M., Zhu, Y., Mehta, R.K., Erraguntla, M., Lawley, M.A., 2019. Static and dynamic work activity classification from a single accelerometer: implications for ergonomic assessment of manual handling tasks. IISE Trans. Occup. Ergon. Hum. Factors 7 (1), 59–68.

United States Bone, Joint Initiative, 2020a. Prevalence of select medical conditions. In: Weinstein, S.I., Yelin, E.H., Watkins-Castillo, S.I. (Eds.), The Burden of Musculoskeletal Diseases in the United States (BMUS), United States Bone and Joint Initiative, Rosemont, IL, fourth ed. URL. https://www.boneandjointburden.org/fourth-edition/ib0/prevalence-select-medical-conditions

United States Bone, Joint Initiative, 2020b. Total economic impact on the US economy. In: Weinstein, S.I., Yelin, E.H., Watkins-Castillo, S.I. (Eds.), The Burden of Musculoskeletal Diseases in the United States (BMUS), United States Bone and Joint Initiative, Rosemont, IL, fourth ed. URL. https://www.boneandjointburden.org/fourth-edition/viiie1/total-economic-impact-us-economy

Karhu, O., Kansi, P., Kuorinka, I., 1977. Correcting working postures in industry: a practical method for analysis. Appl. Ergon. 8 (4), 199–201.

Kehagias, A., Nidelkou, E., Petridis, V., 2006. A dynamic programming segmentation procedure for hydrological and environmental time series. Stoch. Environ. Res. Risk Assess. 20 (1), 77–94.

Kim, S., Nussbaum, M.A., 2014. An evaluation of classification algorithms for manual material handling tasks based on data obtained using wearable technologies. Ergonomics 57 (7), 1040–1051.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Lamooki, S.R., Kang, J., Cavuoto, L.A., Megahed, F.M., Jones-Farmer, L.A., 2021. Personalized and nonparametric framework for detecting changes in gait cycles. IEEE Sensor. J. 21 (17), 19236–19246.

Lavielle, M., 2005. Using penalized contrasts for the change-point problem. Signal Process. 85 (8), 1501–1510.

Li, K., Habre, R., Deng, H., Urman, R., Morrison, J., Gilliland, F.D., Ambite, J.L., Stripelis, D., Chiang, Y.-Y., Lin, Y., Bui, A.A., King, C., Hosseini, A., Vliet, E.V., Sarrafzadeh, M., Eckel, S.P., 2019. Applying multivariate segmentation methods to human activity recognition from wearable sensors' data. JMIR mHealth and uHealth 7 (2), e11201.

Lim, S., D'Souza, C., 2020. A narrative review on contemporary and emerging uses of inertial sensing in occupational ergonomics. Int. J. Ind. Ergon. 76, 102937.

Lowe, B.D., Dempsey, P.G., Jones, E.M., 2019. Ergonomics assessment methods used by ergonomics professionals. Appl. Ergon. 81, 102882.

Maman, Z.S., Yazdi, M.A.A., Cavuoto, L.A., Megahed, F.M., 2017. A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. Appl. Ergon. 65, 515–529.

Maman, Z.S., Chen, Y.-J., Baghdadi, A., Lombardo, S., Cavuoto, L.A., Megahed, F.M., 2020. A data analytic framework for physical fatigue management using wearable sensors. Expert Syst. Appl. 155, 113405.

Mao, W., Fathurrahman, H., Lee, Y., Chang, T., 2020. EEG dataset classification using CNN method. In: Journal of Physics: Conference Series, 1456. IOP Publishing, 012017.

Martindale, C.F., Hoenig, F., Strohrmann, C., Eskofier, B.M., 2017. Smart annotation of cyclic data using hierarchical hidden Markov models. Sensors 17 (10), 2328.

Martindale, C.F., Christlein, V., Klumpp, P., Eskofier, B.M., 2021. Wearables-based multi-task gait and activity segmentation using recurrent neural networks. Neurocomputing 432, 250–261.

McAtamney, L., Corlett, E.N., 1993. RULA: a survey method for the investigation of work-related upper limb disorders. Appl. Ergon. 24 (2), 91–99.

Mokhlespour Esfahani, M.I., Nussbaum, M.A., 2018. A "smart" undershirt for tracking upper body motions: task classification and angle estimation. IEEE Sensor. J. 18 (18), 7650–7658.

Mortazavi, B.J., Pourhomayoun, M., Alsheikh, G., Alshurafa, N., Lee, S.I., Sarrafzadeh, M., 2014. Determining the single best axis for exercise repetition recognition and counting on smartwatches. In: 2014 11th International Conference on Wearable and Implantable Body Sensor Networks, vols. 33–38. IEEE.

National Research Council, 1999. The Changing Nature of Work: Implications for Occupational Analysis. The National Academies Press, Washington, DC, ISBN 978-0-309-17292-9. https://doi.org/10.17226/9600.

National Research Council and the Institute of Medicine, 2001. Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremities. The National Academies Press, Washington, DC, ISBN 978-0-309-07284-7. https://doi.org/10.17226/10032.

Nordander, C., Hansson, G.-Å., Ohlsson, K., Arvidsson, I., Balogh, I., Strömberg, U., Rittner, R., Skerfving, S., 2016. Exposure–response relationships for work-related neck and shoulder musculoskeletal disorders–Analyses of pooled uniform data sets. Appl. Ergon. 55, 70–84.

Panahandeh, G., Mohammadiha, N., Leijon, A., Händel, P., 2013. Continuous hidden Markov model for pedestrian activity classification and gait analysis. IEEE Trans. Instrum. Meas. 62 (5), 1073–1083.

Park, J.W., Kim, D.Y., 2017. Standard time estimation of manual tasks via similarity measure of unequal scale time series. IEEE Trans. Human-Machine Syst. 48 (3), 241–251.

Peppoloni, L., Filippeschi, A., Ruffaldi, E., Avizzano, C., 2016. A novel wearable system for the online assessment of risk for biomechanical load in repetitive efforts. Int. J. Ind. Ergon. 52, 1–11.

Pernek, I., Hummel, K.A., Kokol, P., 2013. Exercise repetition detection for resistance training based on smartphones. Personal Ubiquitous Comput. 17 (4), 771–782.

Porta, M., Kim, S., Pau, M., Nussbaum, M.A., 2021. Classifying diverse manual material handling tasks using a single wearable sensor. Appl. Ergon. 93, 103386.

Putz-Anderson, V., Bernard, B.P., Burt, S.E., Cole, L.L., Fairfield-Estill, C., et al., 1997. Musculoskeletal disorders and workplace factors: a critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremity, and low back, Tech. Rep. DHHS (NIOSH) Publication No. 97B141. National Institute for Occupational Safety and Health. URL. https://www.cdc.gov/niosh/docs/97-141/pdfs/97-141.pdf.

Radwin, R.G., Lin, M.L., 1993. An analytical method for characterizing repetitive motion and postural stress using spectral analysis. Ergonomics 36 (4), 379–389.

Ranavolo, A., Draicchio, F., Varrecchia, T., Silvetti, A., Iavicoli, S., 2018. Wearable monitoring devices for biomechanical risk assessment at work: current status and future challenges—a systematic review. Int. J. Environ. Res. Publ. Health 15 (9), 2001.

Ravi, D., Wong, C., Lo, B., Yang, G.-Z., 2016. Deep learning for human activity recognition: a resource efficient implementation on low-power devices. In: 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), vols. 71–76. IEEE.

Ronao, C.A., Cho, S.-B., 2016. Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst. Appl. 59, 235–244.

Scott, K.A., Browning, R.C., 2016. Occupational physical activity assessment for chronic disease prevention and management: a review of methods for both occupational health practitioners and researchers. J. Occup. Environ. Hyg. 13 (6), 451–463.

Stefana, E., Marciano, F., Rossi, D., Cocca, P., Tomasoni, G., 2021. Wearable devices for ergonomics: a systematic literature review. Sensors 21 (3), 777.

Takala, E.-P., Pehkonen, I., Forsman, M., Hansson, G.-Å., Mathiassen, S.E., Neumann, W. P., Sjøgaard, G., Veiersted, K.B., Westgaard, R.H., Winkel, J., 2010. Systematic evaluation of observational methods assessing biomechanical exposures at work. Scand. J. Work 3–24. Environment & Health.

Tsao, L., Li, L., Ma, L., 2018. Human work and status evaluation based on wearable sensors in human factors and ergonomics: a review. IEEE Trans. Human-Machine Syst. 49 (1), 72–84.

G. J. van den Burg, C. K. Williams, An Evaluation of Change Point Detection Algorithms, arXiv preprint arXiv:2003.06222 .

van der Beek, A.J., Frings-Dresen, M., 1998. Assessment of mechanical exposure in ergonomic epidemiology. Occup. Environ. Med. 55 (5), 291–299.

van Hees, V.T., Golubic, R., Ekelund, U., Brage, S., 2013. Impact of study design on development and evaluation of an activity-type classifier. J. Appl. Physiol. 114 (8), 1042–1051.

Vergara, X., Bhatnagar, M., Fordyce, T., 2020. Exploratory narrative text analysis to characterize tasks associated with injuries among electric utility line workers: EPRI Occupational Health and Safety Database 1995–2013. Am. J. Ind. Med. 64 (3), 198–207.

Wang, T., Lu, C., Sun, Y., Yang, M., Liu, C., Ou, C., 2021. Automatic ECG classification using continuous wavelet transform and convolutional neural network. Entropy 23 (1), 119.

Waters, T.R., Putz-Anderson, V., Garg, A., 1994. Applications manual for the revised NIOSH lifting equation, Tech. Rep. DHHS (NIOSH) Publication No. 94-110. National Institute for Occupational Safety and Health. URL. https://www.cdc.gov/niosh/docs/94-110/pdfs/94-110.pdf.

Wirth, R., Hipp, J., 2000. CRISP-DM: Towards a standard process model for data mining. In: Agrawal, R., Stolorz, P., Piatetsky, G. (Eds.), Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. Springer-Verlag London, UK, pp. 29–39.

Yhdego, H., Li, J., Morrison, S., Audette, M., Paolini, C., Sarkar, M., Okhravi, H., 2019. Towards musculoskeletal simulation-aware fall injury mitigation: transfer learning with deep CNN for fall detection. In: 2019 Spring Simulation Conference (SpringSim), vols. 1–12. IEEE.

Zeng, M., Nguyen, L.T., Yu, B., Mengshoel, O.J., Zhu, J., Wu, P., Zhang, J., 2014. Convolutional neural networks for human activity recognition using mobile sensors. In: 6th International Conference on Mobile Computing, Applications and Services. IEEE, pp. 197–205.