

1 **A Data-Driven Approach to Classifying Manual Material Handling Tasks Using Markerless**  
2 **Motion Capture and Deep Learning Models**

3

4 <sup>a</sup>Aanuoluwapo Ojelade, <https://orcid.org/0000-0001-9715-3254>

5 <sup>b</sup>Mohammad Sadra Rajabi, <https://orcid.org/0000-0002-9100-3973>

6 <sup>b</sup>Sunwook Kim, <https://orcid.org/0000-0003-3624-1781>

7 <sup>b</sup>Maury A. Nussbaum, <https://orcid.org/0000-0002-1887-8431>

8

9 <sup>a</sup>Department of Industrial and Systems Engineering, University at Buffalo, Buffalo NY 14226,  
10 USA

11 <sup>b</sup>Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg VA 24061, USA

12

13

14 Corresponding address: Aanuoluwapo Ojelade

15 Department of Industrial and Systems Engineering,

16 University at Buffalo, 301 Bell Hall, Buffalo, NY 14226, USA.

17 Phone: 7166454721. Email: [aojelade@buffalo.edu](mailto:aojelade@buffalo.edu)

18

19

20 Acknowledgement: all authors have made substantial contributions to all of the following: (1)  
21 the conception and design of the study, or acquisition of data, or analysis and interpretation of  
22 data, (2) drafting the article or revising it critically for important intellectual content, (3) final  
23 approval of the version to be submitted.

24

25

1 **Abstract**

2 Work-related musculoskeletal disorders (WMSDs) are prevalent problems that encompass a  
3 range of conditions affecting muscles, tendons, and nerves due to repetitive strain, non-neutral  
4 postures, and forceful exertions. These disorders lead to pain, reduced productivity and  
5 substantial healthcare costs. Effective physical exposure assessment tools are needed in the  
6 workplace to quantify WMSD risks and the association between exposure and risks. While  
7 several tools are available, they are often limited in scope and lack the ability to assess physical  
8 risks continuously. In this laboratory-based study, we evaluated a data-driven approach to  
9 continuously classify manual material handling tasks and specific task conditions using machine  
10 learning approach, specifically deep learning models. Specifically, kinematic data from  
11 markerless motion capture (MMC) system was used as input for various recurrent neural  
12 networks to classify among eight distinct manual material handling tasks: box lifting,  
13 asymmetric box lifting, box carriage, box pushing, box pulling, cart pushing, overhead lifting,  
14 and box lowering. The models we tested include bidirectional long-short term memory, gated  
15 recurrent units, and bidirectional gated recurrent units. We also classified specific task  
16 conditions, such as hand configurations and initial lifting height. Overall, using the MMC's  
17 kinematic data led to satisfactory results (e.g., accuracy of 80 – 94%) in classifying the tasks and  
18 the task conditions. Our results, though, also emphasize that classification performance varied  
19 across different feature sets, tasks, and between males and females. Nonetheless, use of MMC  
20 demonstrates clear potential for physical exposure assessment.

21

22 Keywords: Physical exposure assessment, Musculoskeletal disorders, Machine learning, sex  
23 differences, Computer vision

24

25

## 1 1.0 Introduction

2 Work-related musculoskeletal disorders (WMSDs) are injuries or dysfunctions affecting bones,  
3 nerves, tendons, muscles, and spinal discs (da Costa & Vieira, 2010). These injuries and  
4 degenerative conditions include nerve compression disorders, soreness, sprains, and strains  
5 (Punnett & Wegman, 2004). WMSDs continue to be an important occupational health concern.  
6 In the United States, WMSDs led to a median of 14 days away from work (U.S. Bureau of Labor  
7 Statistics, 2021) and involved substantial direct costs of about \$2.24 billion annually to  
8 employers (Liberty Mutual Insurance, 2023). Key risk factors contributing to the development of  
9 WMSDs include forceful exertions, repetition, and non-neutral postures. Such risk factors are  
10 highly prevalent during manual material handling (MMH) tasks like lifting, lowering, pushing,  
11 pulling, holding, and carrying (Andersen et al., 2003; Bernard & Putz-Anderson, 1997; da Costa  
12 & Vieira, 2010; Punnett & Wegman, 2004). Effective assessment of physical exposures to such  
13 risks, though, is critical to developing targeted interventions, and more generally for quantifying  
14 associations between exposures and risks or doses (Marras et al., 2009; Plantard et al., 2017;  
15 Waters et al., 2007). A fundamental aspect of physical exposure assessment involves  
16 distinguishing the specific MMH tasks performed (i.e., task classification). Classifying MMH  
17 tasks is crucial for effective risk assessment and developing targeted risk reduction strategies (Li  
18 & Buckle, 1999). Accurately identifying MMH tasks (e.g., lifting, pushing, pulling, and carrying)  
19 enables us to assess the conditions that make each activity potentially harmful to workers. For  
20 instance, a given trunk posture could be risky or not, depending on whether a worker was lifting  
21 a box or pushing a cart. Thus, a complete exposure assessment requires knowing what tasks  
22 people are doing beyond simple kinematics, which have been the focus of much earlier work.  
23 Additionally, accurate task identification can allow for selecting appropriate assessment tools  
24 (e.g., NIOSH Lifting Equation, Liberty Mutual Tables for pushing/pulling), as different tools  
25 apply to different tasks.

26 Physical exposure assessments require methods that can classify MMH tasks accurately and that  
27 are compatible with the work environment. Physical exposure can be assessed using self-reports,  
28 human observations, and direct measurements (David, 2005; Li & Buckle, 1999). Self-report and  
29 human observation approaches are quick, straightforward, and require no advanced technologies.  
30 However, these approaches may be affected by individual biases or sub-optimal workplace  
31 conditions such as occlusions (Pedersen et al., 2016; Plantard et al., 2017). Alternatively, direct  
32 measurements often involve attaching sensors to objects or directly to a worker's body, such as  
33 to obtain postural or force data (David, 2005; Lim & D'Souza, 2020). Examples of such sensors  
34 are inertial measurement units, goniometers, electromyography, and force sensors. Direct  
35 measurements provide objective and precise data for physical exposure assessment. While recent  
36 systems enable real-time data collection and analysis (Lim & D'Souza, 2020; Lind et al., 2023;  
37 Nath et al., 2017), widespread implementation may face challenges, including equipment costs  
38 (Dempsey et al., 2005; Schall Jr et al., 2018), data processing complexity, and potential impacts  
39 on worker comfort (Kent et al., 2015; McNamara et al., 2016; Zhang et al., 2022) and behavior  
40 (e.g., alter movement pattern; Jacobs et al., 2019; le Feber et al., 2021). There is clear value in  
41 automating data collection and analysis to address the challenges posed by existing physical  
42 exposure assessment methods.

43 Recent advancements in sensor technology and deep learning have introduced new alternatives  
44 to perform occupational physical exposure assessments – specifically through computer-based

1 methods (MassirisFernández et al., 2020). Of relevance here are *markerless motion capture*  
2 (MMC) systems, both plain and depth cameras. Existing work demonstrates the feasibility of a  
3 computer-based assessment approach for motion tracking and quantifying physical exposures,  
4 without requiring on-body sensors (Ghezelbash et al., 2024; Han et al., 2013; Plantard et al.,  
5 2017; Roberts et al., 2020; Starbuck et al., 2014; Tang & Golparvar-Fard, 2021; Yun et al., 2025;  
6 Zhang et al., 2018; Zhou et al., 2024). Some example applications emphasize such feasibility. In  
7 one, data extracted from a MMC system were used to automatically detect activities such as  
8 walking and specific tasks during construction drywall installation (Khosrowpour et al., 2014).  
9 MMC data also facilitated classifying different construction worker actions when laying bricks,  
10 transporting rebar, and making formwork (Yang et al., 2016). **Masonry activities—including**  
11 **laying mortar, lifting bricks, placing bricks, and cleaning excess mortar—were automatically**  
12 **identified using data from multiple MMC viewing angles (Yun et al., 2025). Additionally, Han et**  
13 **al. (2013) used MMC data to classify when construction workers ascend and descend ladders.**

14 However, reported applications of MMC have four important limitations. First, it is unclear if  
15 existing task classifiers can be applied when the tasks of interest include relatively complex  
16 motions. In both studies noted above (Khosrowpour et al., 2014; Yang et al., 2016), reasonable  
17 classification accuracy of ~75% was reported for construction tasks, such as shoveling and  
18 transporting, but accuracy substantially decreased for tasks that include similar body motions  
19 (e.g., bolting and plastering). Second, selecting an effective classification algorithm and input  
20 variables remains challenging, since classification performance can depend on the specific  
21 algorithm used and/or the inputs (aka *features*) used. Several machine learning and deep learning  
22 algorithms have been explored – such as Support Vector Machines, K-Nearest Neighbors,  
23 Decision Trees, and Neural Networks (e.g., Escorcía et al., 2012; Park et al., 2016; Song et al.,  
24 2010; Yu et al., 2019; Zhan et al., 2012; Zhang & Tian, 2012) – and performance on a given task  
25 clearly varies based on the specific algorithm and input variables employed. Third, the tasks  
26 included in earlier reports have often lack adequate representation of occupational tasks  
27 generally and MMH tasks specifically. Fourth, current computer-based assessments mainly  
28 classify discrete tasks, and it is unclear if they can accurately capture and analyze continuous or  
29 complex task sequences. This limitation is critical because discrete analysis can underestimate  
30 physical exposures in MMH tasks, highlighting the need for continuous quantification.

31 Our purpose in this study was thus to investigate the performance of an MMC system, together  
32 with deep learning algorithms, for classifying diverse MMH tasks during a simulated complex  
33 job. Specifically, we explored the relative performance of using different deep learning  
34 algorithms as an ergonomic exposure assessment approach for identifying specific MMH tasks  
35 and for distinguishing among different task conditions (e.g., initial lifting height). Several deep  
36 learning algorithms were tested, since no single model was expected to be best suited for MMH  
37 task classification (Jozefowicz et al., 2015). We examined the performance of recurrent neural  
38 networks (RNN) in analyzing sequential MMH tasks. RNN models were selected over other  
39 classification algorithms (e.g., artificial neural networks) for their superior ability to process  
40 sequential data and preserve contextual relationships within time series data (Arisoy et al., 2015;  
41 Logar et al., 1993; Schuster & Paliwal, 1997). Additionally, we evaluated the effects of various  
42 input variables (i.e., feature sets derived from kinematic data) on MMH task classification  
43 performance. Our study was exploratory in nature, seeking results that could inform future  
44 assessments of physical exposures using MMC. We expected the performance of deep learning

1 algorithms to depend on the feature sets used, and to differ between specific MMH tasks and task  
2 conditions, and with biological sex. The latter expectation was based on evidence of kinematic  
3 differences in how males and females perform MMH tasks (Martinez et al., 2019; Plamondon et  
4 al., 2017).

## 5 **2.0 Methods**

6 MMH tasks were simulated in a controlled, laboratory setting, and these tasks were then  
7 classified using body kinematics obtained from an MMC system. Diverse tasks were simulated,  
8 representative of physically-demanding activities in several occupational sectors (e.g., lifting,  
9 carrying, pushing). Several task classifiers were explored using different deep learning methods  
10 and feature sets, and the performance of these classifiers were examined using common metrics.

### 11 **2.1 Participants**

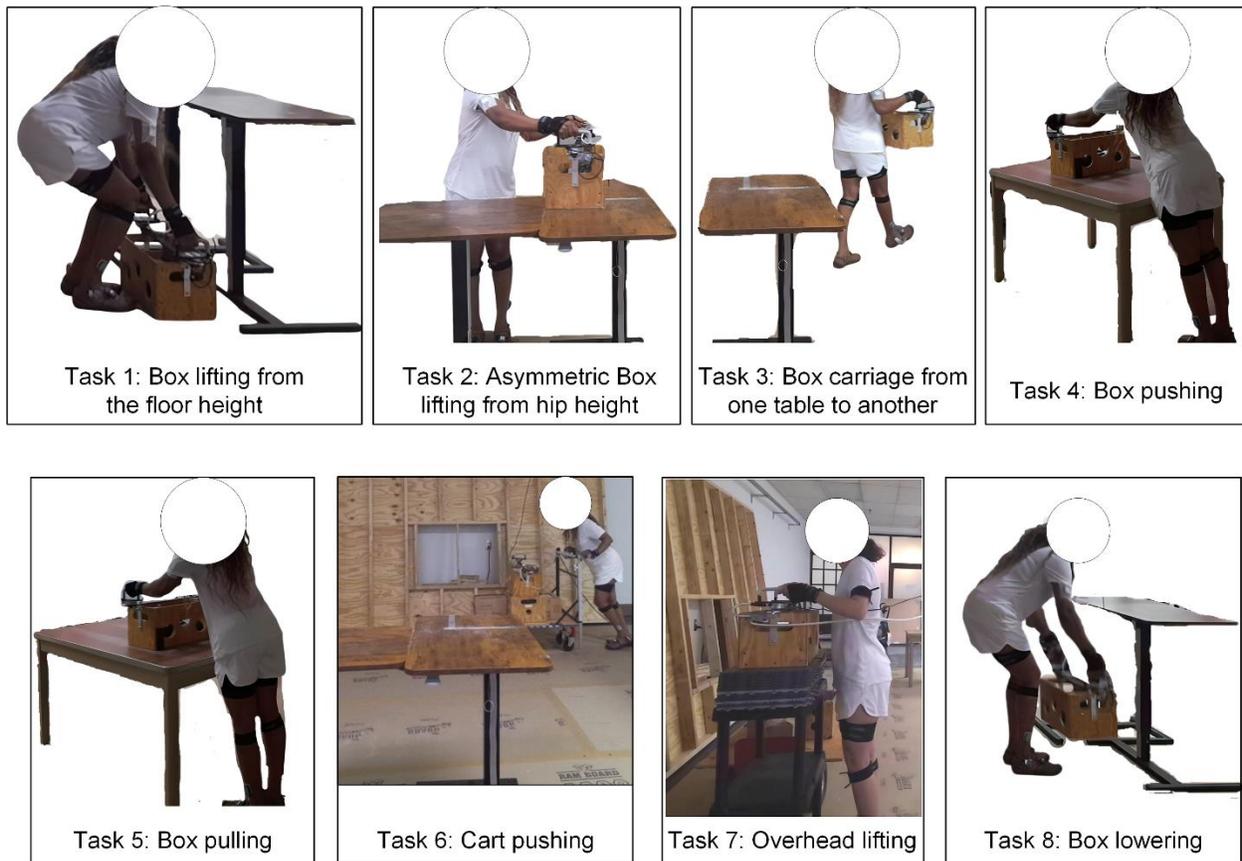
12 A convenience sample of 36 young (22 males and 14 females) participants completed the study  
13 and were recruited from the university and local community. Respective means (SD) of age,  
14 body mass, and stature were 27 (4.4) years, 77.5 (12.2) kg, and 176.4 (6.9) cm for the males; and  
15 27 (5.2) years, 68 (12.3) kg, and 170.1 (7.2) cm for the females. Eligibility criteria required  
16 participants to be aged 18-60, with no history of low back injury or musculoskeletal disorders in  
17 the past 12 months, and to exercise at least twice per week. These criteria were self-reported by  
18 the participants. The research reported herein complied with the tenets of the Declaration of  
19 Helsinki, and the study protocol was approved by the Institutional Review Board at Virginia  
20 Tech. Informed consent was obtained from all participants prior to any data collection.

### 21 **2.2 Task Simulations**

22 Eight MMH tasks were simulated in the laboratory, and these tasks involved some variations of  
23 manual box lifting, carrying, pushing, pulling, and reaching. The tasks simulated here are similar  
24 to those used in an earlier study that evaluated the efficacy of several classification models for  
25 MMH tasks (Kim and Nussbaum (2014)). Specifically, six within the current set of tasks were  
26 also included in the noted earlier study. Distinct here, however, was the inclusion of cart pushing  
27 and the use of different hand configurations, box masses, lift origins, and starting positions. The  
28 specific MMH tasks simulated are described below (see also Figures 1 and 3). A single wood box  
29 (width = 26.0 cm; depth = 41.0 cm; and height = 23.5 cm) was used across all tasks.

- 30 • Task 1: symmetric box lifting from two different origins (floor and individual knee  
31 height) to an individual hip height. Hip height was defined as the vertical distance from  
32 the floor to the greater trochanter.
- 33 • Task 2: asymmetric box lifting, from a table placed in front of the participant to another  
34 positioned 90° to the left of the participant. The two tables were adjusted to individual hip  
35 height.
- 36 • Task 3: box carriage from one table to another (i.e., lifting from hip height, carrying, and  
37 lowering to a height of 0.74m). Participants carried the box over a distance of 2.4 m,  
38 selected as the 50<sup>th</sup> percentile of the carrying distance of the U.S. workforce (Ciriello et al.,  
39 1999).
- 40 • Task 4: box pushing over a distance of ~0.7 m, with table height = 0.74 m.
- 41 • Task 5: box pulling toward the body over a distance of ~0.7m, with the table height fixed  
42 at 0.74 m.

- 1 • Task 6: cart pushing at individual waist height over a distance of ~1.8m. Cart mass was
- 2 fixed at 86 kg, including the box, representing common cart loads (Hoozemans et al.,
- 3 2004). Cart pushing was completed only using two hands.
- 4 • Task 7: overhead lifting from cart height to individual overhead height, with the box
- 5 lifted from a height of 0.56 m (height of the cart). Overhead height was defined from
- 6 individual anthropometric measures as the distance between the lateral epicondyle and
- 7 the floor when the shoulder was flexed at 80°. This anthropometric measure was selected
- 8 since working repeatedly with arm flexion or abduction beyond 80° has been associated
- 9 with shoulder disorders (Bernard & Putz-Anderson, 1997).
- 10 • Task 8: lowering to different origins (floor and individual knee height) from overhead
- 11 height. Participants carried the box over a distance of ~1.3m.



12  
13 Figure 1: Illustrations of the simulated manual material handling tasks.

14 We used a repeated-measures design, in which each participant completed a total of 36 scenarios

15 were completed, involving all possible combinations of the task conditions: a) three levels of

16 *Hand Configuration*; b) three levels of *Box Mass*; c) two levels of *Lift Origin*; and d) two levels

17 of *Start Position*. *Hand Configuration* had three levels: broad, narrow, and one-hand (Figure 2).

18 Both the broad and narrow hand configurations involved using both hands, for which the handles

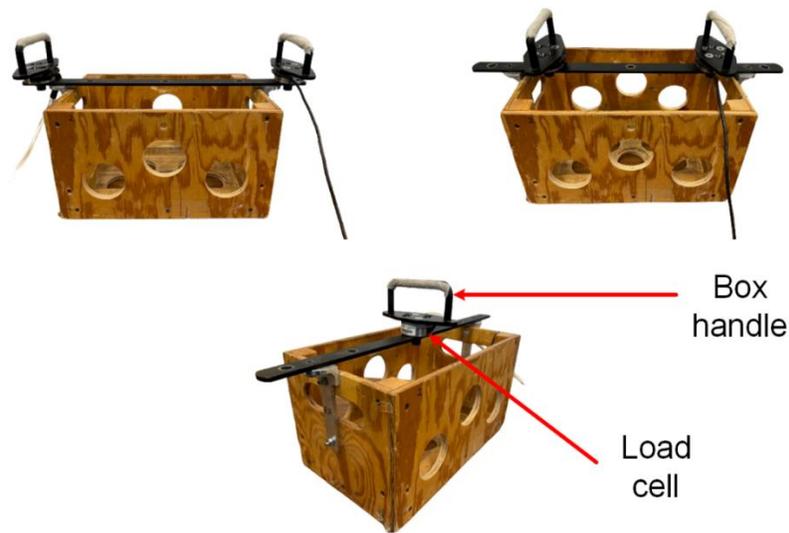
19 were spaced at 52 and 33 cm apart, respectively. In the one-hand configuration, the box handle

20 was positioned in the middle of the box. Multiple hand configurations were used here since

21 different hand widths and one- vs. two-handed lifting methods impose different biomechanical

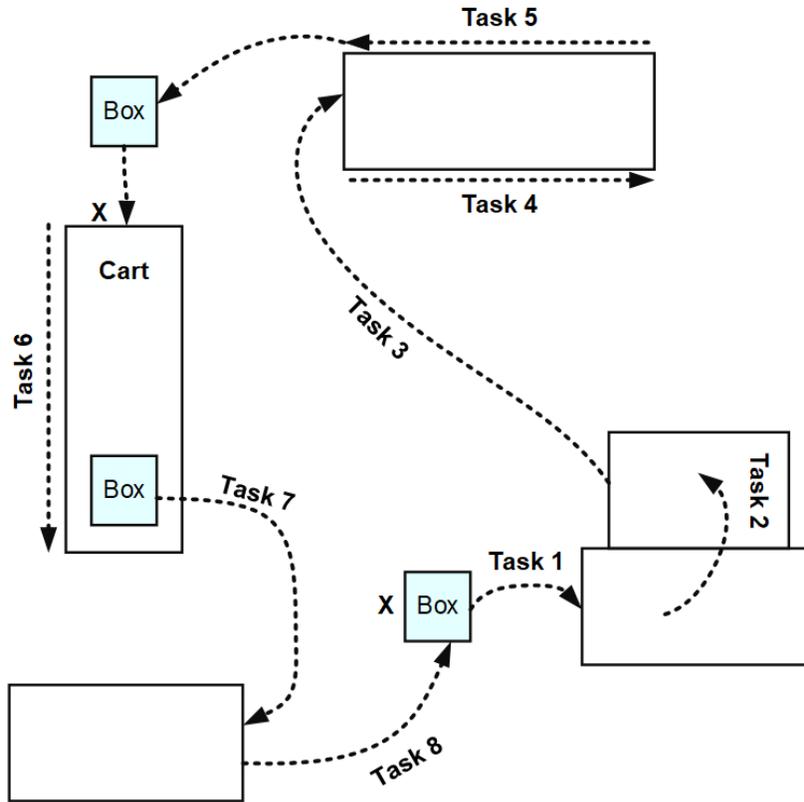
22 demands on the lower back (Garg et al., 1982; Gary et al., 1996; Marras & Davis, 1998). Three

1 levels of *Box Mass* – 6, 9, and 12 kg – were used for the broad and narrow hand configuration,  
2 while masses of 5, 7, and 9 kg were used for one-hand hand configuration. Different box masses  
3 for one vs. two hands lifting tasks were selected, based on results from pilot testing (i.e., to  
4 ensure that most participants could complete all tasks). Specific masses used here were roughly  
5 within the 8<sup>th</sup> and 22<sup>nd</sup> percentiles of masses lifted by the U.S. workforce (Ciriello et al., 1999).  
6 Two levels of *Lift Origin* – floor and individual knee height – were included to impose different  
7 physical exposures during lifting. Finally, two levels of *Starting Position* were used to impose  
8 more task variability. Specifically, one starting position was set at Task 1 and the other at Task 6  
9 (Figure 3).



10

11 Figure 2: Illustration of the three levels of hand configuration: broad (top-left), narrow (top-  
12 right), and one-hand (bottom-middle). Note that the handles were oriented parallel to the short  
13 sides of the box (they appear at an angle in the figure only as an artifact of the lens setting used).



1  
2 Figure 3: Top-view schematic of the simulated tasks. Dotted lines indicate the movement path,  
3 and “X” indicates two alternative starting positions. Descriptions of the eight tasks are provided  
4 in the text.

### 5 2.3 Experimental Procedures

6 Participants completed one experimental session (~3 hrs.), which consisted of training and  
7 experimental phases. During the training phase, participants were introduced to the MMH tasks  
8 and the different task conditions, then practiced the simulated tasks. They were asked to perform  
9 all tasks using their own comfortable work strategies and speed, while assuming they were  
10 working in an industrial environment.

11 In the experimental phase, participants completed multiple trials of the MMH tasks, performing  
12 one trial for each of the 36 task conditions. A study *trial* involved completing all of the eight  
13 MMH tasks sequentially, with a given box mass, hand configuration, and lift origin, and from a  
14 given starting position. The presentation order of *Hand Configuration* was counterbalanced using  
15  $3 \times 3$  balanced Latin Squares. Within a given *Hand Configuration*, the presentation order of *Box*  
16 *Mass* was also counterbalanced using  $3 \times 3$  balanced Latin Squares, whereas the presentation  
17 orders of *Starting Position* and *Lift Origin* were alternated across participants. To mitigate  
18 physical fatigue, a minimum of four minutes of rest was given between each *Hand Position*  
19 condition.

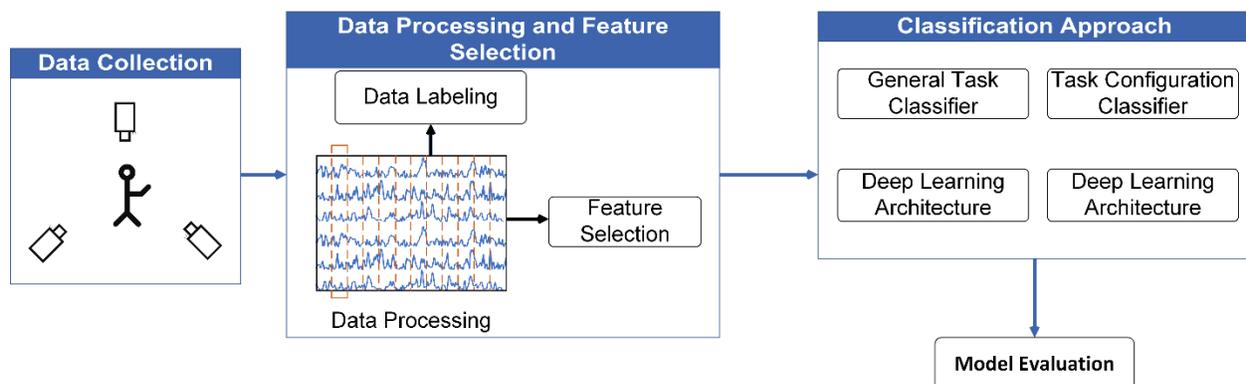
### 20 2.4 Instrumentation

21 Whole-body kinematics were monitored at 30 Hz using three markerless camera systems (Azure  
22 Kinect™, Microsoft Corporation, Seattle WA, USA). Azure Kinect, the latest depth camera from

1 Microsoft, was selected over other system because it has a relatively high resolution and a global  
 2 shutter (which minimizes motion artefacts); we thus expected this system would provide  
 3 sufficient data quality and could be useful even in dynamic outdoor environments. These systems  
 4 were positioned ~1.74 m from the edge of the work area, a configuration that was determined to  
 5 be effective during pilot testing and that aimed to optimize the coverage of the narrow camera  
 6 field of view (see Figure Appendix.1 or A.1). The three Azure Kinects were time-synchronized  
 7 using a 3.5 mm auxiliary cable connected in a daisy-chain configuration, where one Azure  
 8 Kinect was designated as the primary device, with the remaining two as secondary devices. An  
 9 iPi Recorder (iPi Soft®; [www.ipisoft.com](http://www.ipisoft.com)) was used for sampling from the Azure Kinects. At the  
 10 start of each trial, participants completed a calibration step by standing in a “T-pose” while  
 11 facing one of the markerless cameras. This required them standing upright, with their arms  
 12 abducted horizontally, and their feet together and pointing forward. This calibration was used to  
 13 establish a clear and consistent starting point for tracking body joints (see:  
 14 [https://docs.ipisoft.com/User\\_Guide\\_for\\_Multiple\\_Depth\\_Sensors\\_Configuration#T-Pose](https://docs.ipisoft.com/User_Guide_for_Multiple_Depth_Sensors_Configuration#T-Pose)).

## 15 2.5 Data Processing and Feature Selection

16 Task classification can be viewed as multi-staged processing following data collection: data  
 17 processing, feature selection, data labeling, classification approach, and model evaluation (Figure  
 18 4). Each of these stages is discussed in detail subsequently.



19

20

Figure 4: Overview of the task classification process.

### 21 2.5.1 Data Processing

22 Data recorded from the markerless camera systems were processed using iPi Motion Capture  
 23 Studio (version 4.6.3, iPi Soft ®; [www.ipisoft.com](http://www.ipisoft.com)). Using this software, 3D body motions were  
 24 tracked, and key body joints were identified from the video recorded using the iPi Recorder.  
 25 Seventeen body segments were tracked and stored for further processing – pelvis; lower, middle,  
 26 and upper spine; neck and bilateral acromioclavicular; shoulders, elbows, hips, knees, and  
 27 taluses. Tri-axial positions and quaternion joint rotations were extracted for each of the body  
 28 joints using the iPi Biomechanical add-on software. Joint kinematics were low-pass filtered (6  
 29 Hz cutoff; 4<sup>th</sup> order Butterworth; bidirectional) to remove sensor noise and other artifacts, with  
 30 the cutoff frequency determined using residual analysis (Winter, 2009). Filtered joint kinematics  
 31 were then normalized, using the Min-Max-Scaler function (Pedregosa et al., 2011), to a (0, 1)  
 32 range across participants, given that deep learning models are sensitive to unscaled data  
 33 (Djordjević et al., 2022; LeCun et al., 2002; Singh & Singh, 2020). All subsequent offline data  
 34 processing was completed using Python (ver. 3.10.11; [www.python.org](http://www.python.org)).

## 1 **2.5.2 Feature Selection**

2 Features are independent measurable properties or characteristics of the data that serve as input  
3 for deep learning models. In their basic form, features can be raw data. In some instances,  
4 however, the most informative features are selected while discarding irrelevant or redundant  
5 ones, by using feature selection algorithms (Markovitch & Rosenstein, 2002). Therefore, we  
6 considered two types of features: raw and “informative” features. First, the processed joint  
7 kinematics, which included 119 features from the processed kinematics (i.e., raw features = RF),  
8 consisting of tri-axial positions and quaternion elements of the 17 body joints, were included as  
9 input for model training described below. Second, high-dimensional features may lead to longer  
10 classification processes and overfitting. Reducing dimensionality, by selecting the most  
11 informative features, can simplify deep learning models and help improve model performance  
12 and interpretability (Chen et al., 2017). Dimensionality reduction is especially valuable in the  
13 context of MMC systems, since occlusions from the body or the environment can lead to data  
14 loss and inaccurate pose estimations (Plantard et al., 2017). A filter-based method for feature  
15 selection – Minimal-Redundancy-Maximum-Relevancy (Peng et al., 2005) – was used to select  
16 subsets of features from among the entire kinematic feature pool or RF. The primary goal of this  
17 approach is to find a feature subset that minimizes redundancy between features, while  
18 maximizing their relevance to the target variable, and more details on this method have been  
19 reported elsewhere (Peng et al., 2005). Since there is no consensus on the number of features that  
20 could yield the best model performance, we arbitrarily tested the top 60 (TOP-60) and top 80  
21 (TOP-80) features as input for each deep learning model. Listings of these two feature sets for  
22 each classification scheme (see Section 2.6.3) are provided in Tables A.1 – A.6.

## 23 **2.5.3 Data Labeling**

24 Processed data were labeled manually for the specific tasks performed by visually observing the  
25 recorded RGB-D data (served as the ground truth). In each MMH task that involved  
26 manipulating a box, the task began when participants touched the box handle and ended when  
27 they removed their hands from the box. In the case of cart pushing (Task 6), the task started  
28 when participants touched the cart handles and ended when they removed their hands from the  
29 box. Data corresponding to when participants were idle between each task were removed, and  
30 the remaining data were concatenated (see example in Figure A.3).

## 31 **2.6 Classification Approach**

### 32 **2.6.1 Deep Learning Algorithms**

33 Recurrent neural network (RNN) architectures were used to classify MMH tasks using MMC-  
34 derived kinematic measures. RNNs, which are a variant of an artificial deep learning network,  
35 are designed to process and recognize patterns in sequential data (Levin, 1990). Unlike other  
36 neural networks and deep learning models that process data in a cross-sectional way, RNNs  
37 maintain contextual relationships over time series segments. Three RNN models were used and  
38 compared here for classifying the MMH tasks: 1) Bidirectional Long Short-Term Memory (Bi-  
39 LSTM); 2) Gated Recurrent Units (GRU); and 3) Bidirectional Gated Recurrent Units (BGRU).  
40 Bi-LSTM, which is a variant of LSTM, incorporates both forward and backward LSTM models  
41 to process data in both directions, effectively addressing the vanishing gradient problem  
42 (Hochreiter & Schmidhuber, 1997). This problem occurs when the gradients of the loss function  
43 become extremely small, causing the network weights to update very slowly (Hochreiter &

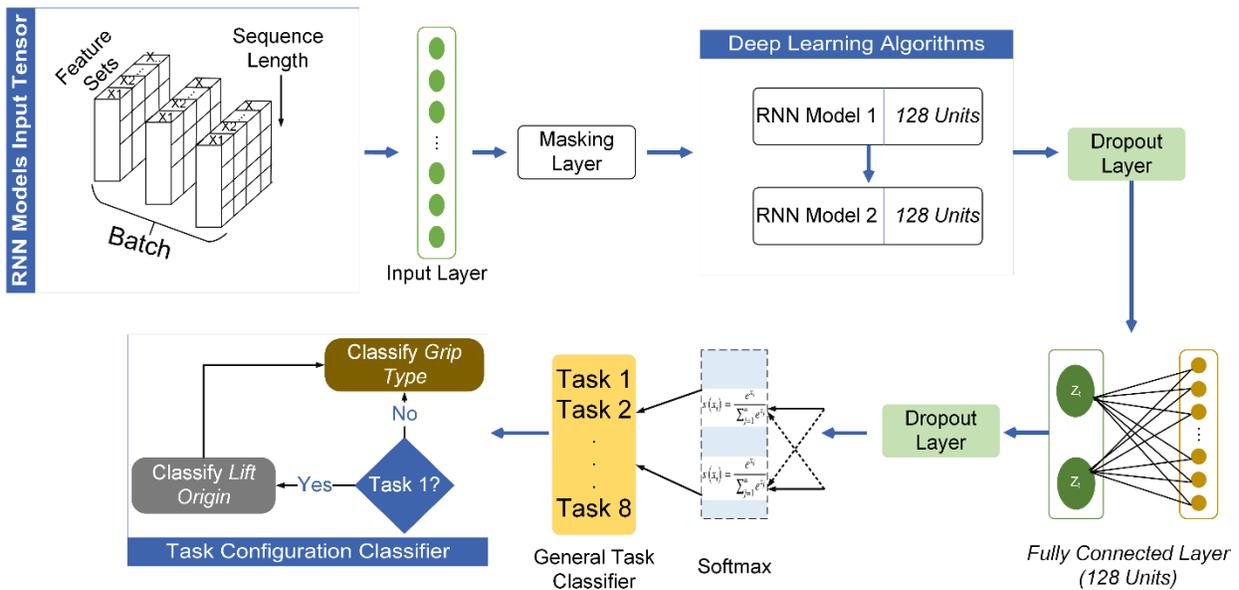
1 Schmidhuber, 1997; Jozefowicz et al., 2015). The gating mechanisms of a Bi-LSTM includes  
 2 three gates (i.e., input, forget, and output) and allows for modifications of Bi-LSTM cell states.  
 3 By integrating inputs from past-to-future and future-to-past directions, this model enhances task  
 4 classification performance (Graves et al., 2013; Yang et al., 2020; Zhou et al., 2022). A GRU  
 5 model simplifies the LSTM model by using fewer parameters, improving its efficiency in  
 6 *understanding* long-term dependencies. Finally, a BGRU is built upon the GRU model by adding  
 7 a bi-directional layer, providing the model output layer with complete contextual information of  
 8 the input data at each time point. In brief, the input data are passed through feedforward and  
 9 backward GRU networks, and the outputs of these two pathways are connected at the same outer  
 10 layer. Some studies have shown that BGRU models are suitable for classification problems, such  
 11 as for human identification (Lynn et al., 2019) and for dialog intent classification (Wang et al.,  
 12 2020).

13

14 **2.6.2 Model Architecture**

15 The model architecture consisted of an input layer, an RNN model layer (i.e., Bi-LSTM, GRU,  
 16 or BGRU), a masking layer, dropout functions, dense layers, and an output layer (Figure 5).  
 17 Keras, a high-level deep learning library, was used to implement the model architecture since it  
 18 enabled the efficient implementation of RNN model layers, masking layers, and model  
 19 configurations (Chollet & Others, 2015).

20



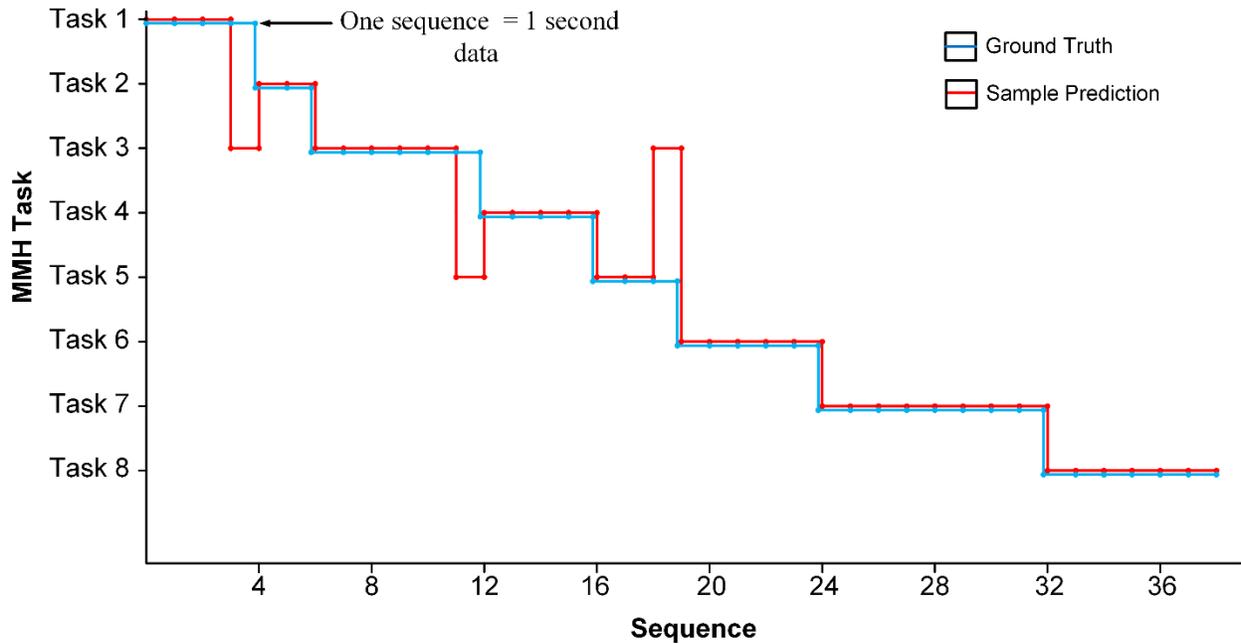
21

22 Figure 5: Overall architecture of recurrent neural network models for classification.

23 The input layer of each RNN model was designed to take a multidimensional matrix defined by  
 24 sequence length, feature dimension, and batch size. Sequence refers to the ordered set of input  
 25 data points processed sequentially over time steps, which was fixed at 30 here (Figure A.2),  
 26 corresponding to 1 second of data at the 30 Hz sampling rate. This sequence length was selected  
 27 to account for tasks that have short duration, and it has been used in previous human activity

1 recognition studies to improve recognition of short duration tasks (Bulling et al., 2014; Capela et  
 2 al., 2015). To account for the varying duration of each task, the sequence was zero-padded  
 3 (Figure A.2, Appendix A) to ensure a consistent length (e.g., see (Dwarampudi & Reddy, 2019)).  
 4 A masking layer was added to each of the RNN models to mitigate padding effects. Specifically,  
 5 this layer skips any sequences that have the special masking value. Feature dimension is the  
 6 number of features, based on the dimensions of the features generated earlier. Raw, TOP-60, and  
 7 TOP-80 features were 119-, 60- and 80-dimensional vectors, respectively. Finally, the batch size  
 8 was set to 64, which was determined as the optimal size during initial testing. The input layer  
 9 was connected to each of the RNN model architectures, consisting of a dropout function, a dense  
 10 layer, an optimizer, a loss function, and an output classification layer. Our output classification  
 11 layer used a many-to-one architecture, wherein a single output is synthesized from the input data.

12 Classification decisions were made *sequence-to-sequence*, specifically one decision for each  
 13 second of data. Each input sequence was classified independently, with the entire sequence being  
 14 considered at every time step (Figure 6). In this context, a time step corresponds to each discrete  
 15 unit of time within an input data sequence, and here the discrete time unit was 1 second. The  
 16 number of sequences varied across tasks due to differences in task completion time and the way  
 17 sequences were defined. Using time steps enables RNNs to capture temporal dependencies and  
 18 facilitates modeling of dynamic input sequences. Time steps also enable RNN models to  
 19 maintain contextual relationships between past and present information, making RNNs well-  
 20 suited for tasks involving time series data.



21  
 22 Figure 6: An example of continuous classification using a sequence-to-sequence approach. A Bi-  
 23 LSTM model and the raw feature set were used in this example. Each dot represents a sequence,  
 24 blue lines indicate ground truth, and red lines indicate output from the classification model. In  
 25 this example, three classification errors are evident.

26 **2.6.3 Classification Scheme**

1 Two categories of RNN models were developed – *general task classifiers* and *task configuration*  
2 *classifiers*. General task classifiers were trained using each of the three feature sets (RF, TOP-60,  
3 and TOP-80) to classify the eight simulated MMH tasks. Task configuration classifiers, in  
4 contrast, classified the different configurations within relevant simulated tasks (i.e., Lift Origins  
5 and/or Hand Configuration). We used a multi-stage classification approach using two  
6 classification stages because using a single classifier for classifying both tasks and task  
7 configurations could result in more frequent false positives (Senator, 2005).

#### 8 **2.6.4 Model Training, Validation, Hyperparameters, and Evaluation**

9 For each category of RNN model, experimental data were used as input to the models. We used a  
10 leave-one-subject-out approach to train and validate the RNN models; this approach is a special  
11 case of cross-validation, in which each subject is considered as a “fold”. Thus, data from 35  
12 participants were used for training, with the remaining participant’s data used for validation. This  
13 process was repeated 36 times (i.e., 36-fold cross-validation). While the leave-one-subject-out  
14 method has been used in past work (e.g., Kim & Nussbaum, 2014; Porta et al., 2021), it often  
15 results in high variance in accuracy since participants can perform the same tasks in different  
16 ways (Jordao et al., 2018). This method, however, replicates real-world training and testing,  
17 wherein the model is trained offline using known subject data and then tested on an unseen  
18 subject (Jordao et al., 2018).

19 RNN models have several hyperparameters that are used to control the learning process and  
20 model complexity (Probst et al., 2019). Hyperparameter values were determined here using an  
21 empirical tuning process involving manual adjustments to each parameter and subsequent  
22 evaluation of the resulting model performance. Specific values tested and the final values  
23 adopted are presented in Table A.7. Performance of the RNN models was assessed using four  
24 common metrics – macro accuracy (accuracy), Precision, recall, and F1-score – which are  
25 provided in Figure A.4. Macro accuracy was assessed to account for the class imbalance across  
26 MMH tasks. Accuracy measures the ratio of correct predictions to total predictions. Precision  
27 represents the correctness of positive predictions, while recall captures how well the model  
28 identifies all positive instances. The F1-score provides a balanced metric combining precision  
29 and recall in a single value.

#### 30 **2.7 Statistical Analyses**

31 To account for the differing number of independent variables that could affect each performance  
32 metric, two sets of analyses of variance (ANOVAs) models were used. First, a two-way repeated-  
33 measures ANOVA was used to assess the effects of *Feature set* and *RNN model* on accuracy. The  
34 latter effect had three levels, to represent general task, hand configuration, and lift origin  
35 classifiers. This initial analysis was performed, since accuracy could only be computed across all  
36 eight simulated MMH tasks. Second, separate three-way repeated-measures ANOVAs were used  
37 for the remaining performance metrics (precision, recall, and F1-score). For general task  
38 classification, the independent variables were *Feature set*, *RNN model*, and *MMH task*. Two sets  
39 of models were used for the task configuration classifiers, with different independent variables:  
40 a) *Feature set*, *RNN models*, and *Hand Configuration*; and b) *Feature set*, *RNN models*, and *Lift*  
41 *Origin*. For each ANOVA model, biological sex (*Sex*) was included as a blocking effect,  
42 significant interaction effects were explored using simple-effects testing, and *post hoc* paired  
43 comparisons were completed using the Tukey’s HSD procedure. All statistical analyses were  
44 performed with JMP Pro 16 (SAS, Cary, NC) using the restricted maximum likelihood (REML)

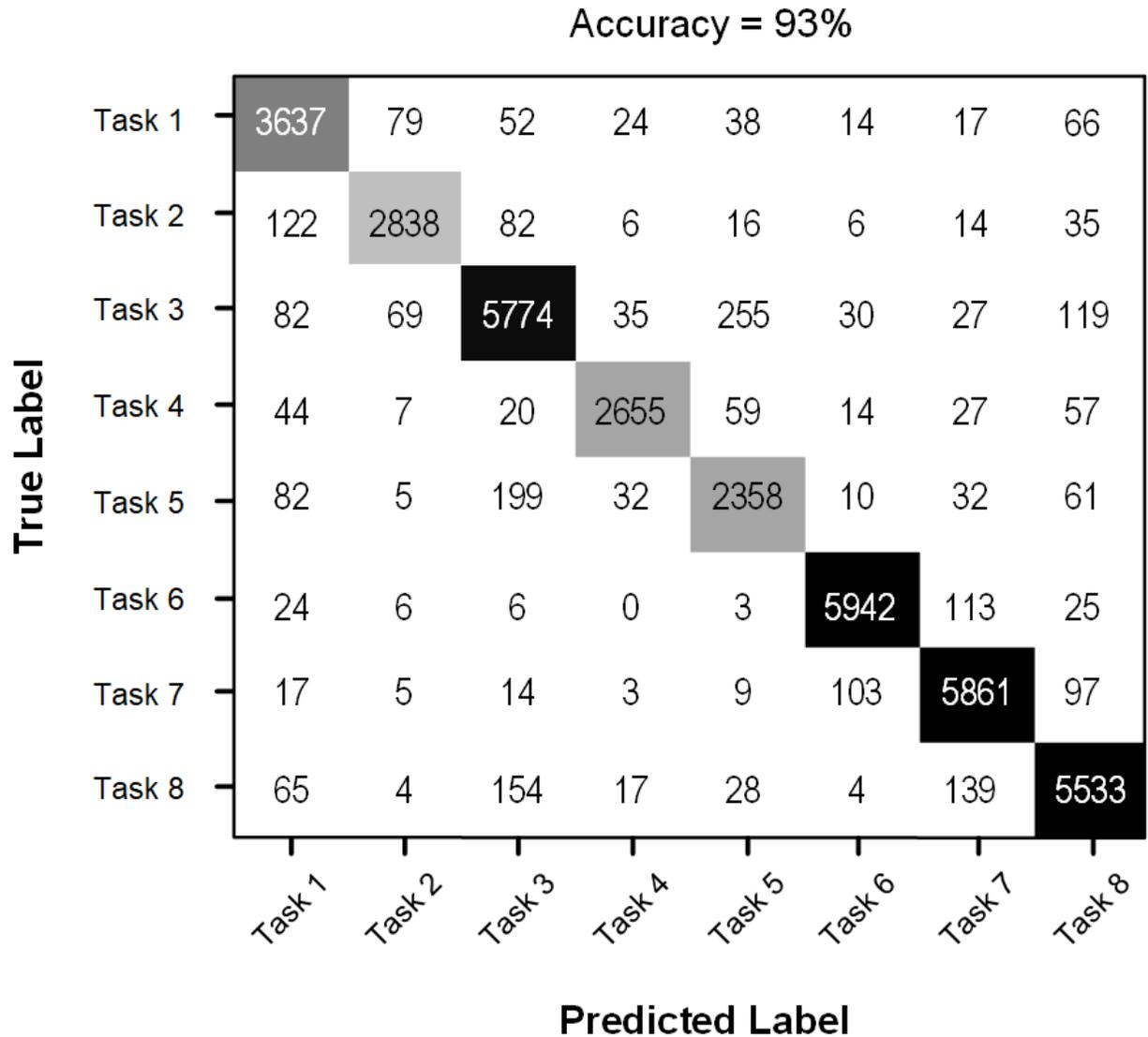
1 method. Parametric model assumptions were verified, and statistical significance was determined  
2 when  $p < 0.05$ . Summary data are reported as least-square means (with 95% confidence  
3 intervals) based on the statistical model fits.

### 4 **3.0 Results**

5 ANOVA results are summarized in Tables A.8 – A.11, and Figures A.5 – A.7 provides confusion  
6 matrices for each category of RNN model. Sample confusion matrices are shown below (Figures  
7 6, 8 and 10) using results representing the best classification performance we obtained. There  
8 were significant main or interactive effects of *RNN model* and *Feature set* for all classification  
9 performance metrics. More detailed results are provided below.

### 10 **3.1 MMH Task Classification Performance**

11 Accuracy: There were significant main effects of both *RNN model* and *Feature set* on accuracy  
12 (Table A.8). The GRU model yielded significantly smaller accuracy (~91%), vs. the BGRU and  
13 Bi-LSTM models (~92%), though the magnitude of the difference was clearly quite small. Using  
14 the TOP-60 feature set led to significantly lower mean GRU accuracy compared to the TOP-80  
15 and RF feature sets, though again the difference was rather small (~90 vs. ~93%, respectively). A  
16 sample confusion matrix is shown in Figure 7; in this case, the model performed well on most  
17 tasks, with an overall accuracy of ~93%.

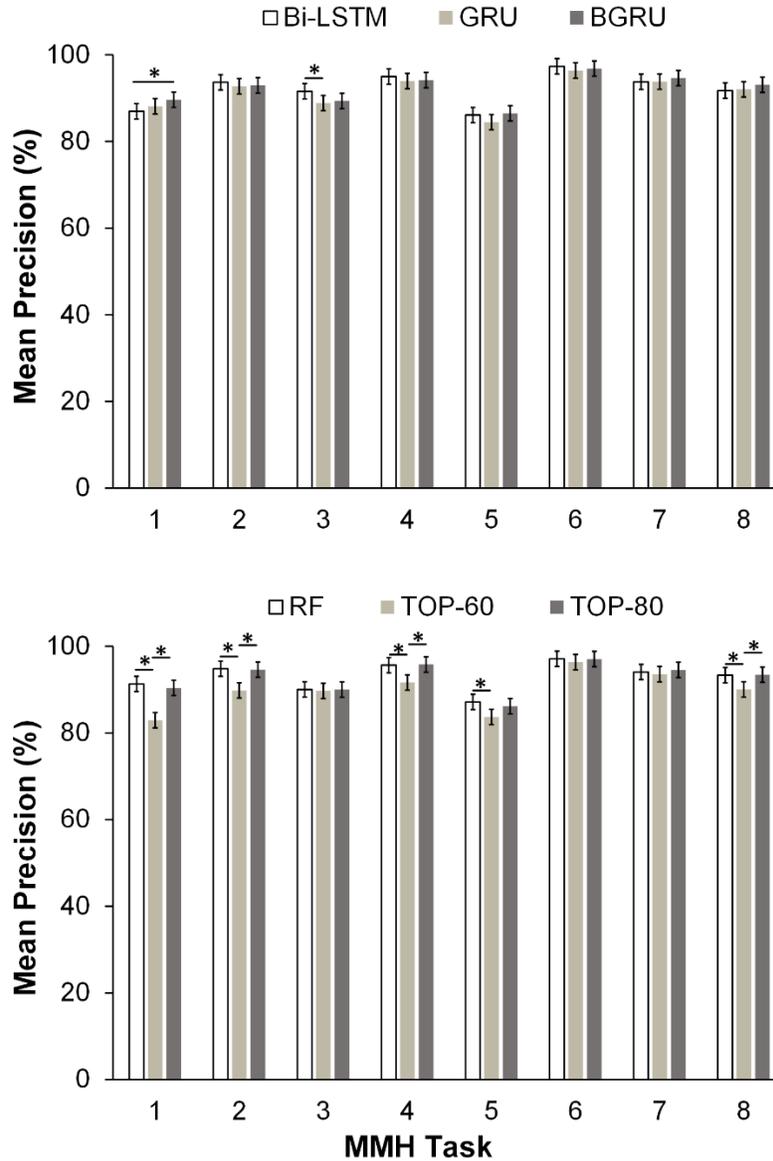


1

2 Figure 7: Overall confusion matrix from using a Bi-LSTM model using the TOP-80 feature set,  
 3 representing the best performance when classifying the MMH tasks. For this and other confusion  
 4 matrices, cells on the main diagonal indicate correct classifications, and lighter shades indicate  
 5 more misclassifications.

6 Precision: *RNN model*, *MMH task*, and *Feature set* main effects, and *RNN model* × *MMH task*,  
 7 *Feature set* × *MMH task*, *Sex* × *MMH task* interaction effects, were all significant (Table A.9).  
 8 Precision was relatively high and comparable between the different RNN models for most of the  
 9 MMH tasks (i.e., ~87-97%; Figure 8). However, using the GRU model led to precision that was  
 10 up to ~3% lower compared to the Bi-LSTM and BGRU models (Figure 8). Simple effects were  
 11 significant, except for the effect of *RNN model* in Tasks 2 ( $p = 0.37$ ), 4 ( $p = 0.29$ ), 6 ( $p = 0.37$ ),  
 12 and 7 ( $p = 0.36$ ). Using the TOP-60 feature set consistently resulted in 2-8% lower precision than  
 13 when using the TOP-80 and RF feature sets, with this difference depending on the specific MMH  
 14 task (Figure 8). Simple effects were significant, except for the effect of *Feature set* in Tasks 3 ( $p$   
 15 = 0.87), 6 ( $p = 0.46$ ), and 7 ( $p = 0.36$ ). Precision in some tasks was ~4% smaller among males

1 than females, though no significant paired differences between sexes were found for any tasks  
 2 (Figure A.8). Simple effect analysis showed that precision differed significantly across MMH  
 3 tasks within each sex ( $p < 0.0001$  for both males and females), indicating the specific MMH task  
 4 performed influenced precision for both sexes.



5  
 6 Figure 8: Interaction effects of *RNN Model* × *MMH task* (top) and *Feature set* × *MMH task*  
 7 (bottom) on classification *precision*. For this and other figures below, error bars indicate 95%  
 8 confidence intervals, and the symbol \* indicates a significant difference between pairs of means.

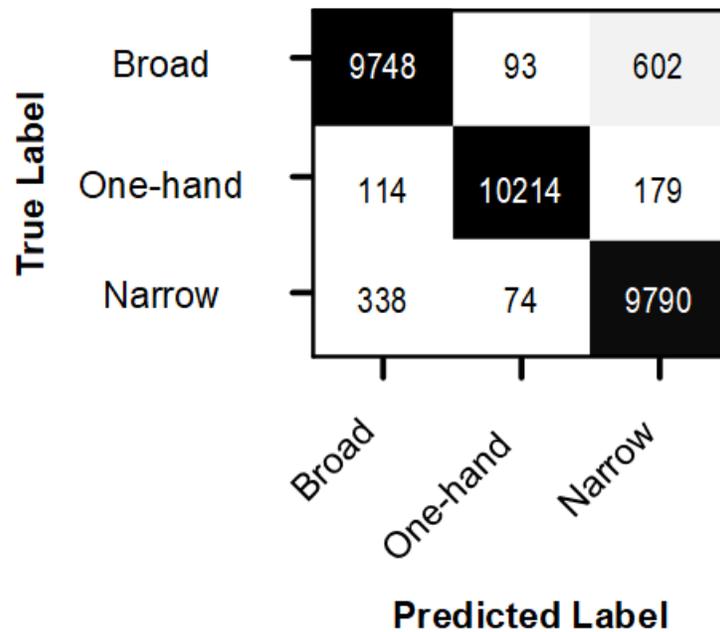
9 Recall and F1-score: *RNN model*, *MMH task*, and *Feature set* main effects, and *Feature set* ×  
 10 *MMH task* and *Sex* × *MMH task* interaction effects, were all significant for both recall and F1-  
 11 score (Table A.9; Figures A.9 and A.10). One consistent observation was that using the TOP-60  
 12 features led to significantly reduced recall and F1-scores compared to TOP-80 and RF, by up to  
 13 7% depending on the specific MMH task (Figures A.9 and A.10). All simple effects were

1 significant, except for the effect of *Feature set* in Tasks 6 ( $p = 0.51$ ) and 7 ( $p = 0.20$ ). No  
 2 significant paired differences were observed between sexes or between MMH tasks. As was the  
 3 case for precision, though, these metrics were somewhat lower among males, by ~1–3%  
 4 depending on the specific MMH task. Simple effect analysis showed that recall and F1-score  
 5 differed significantly across MMH tasks within each sex ( $p < 0.0001$  for both males and  
 6 females).

### 7 **3.2 Classifying Hand Configuration**

8 Accuracy: There was a significant main effect of *Feature set* on accuracy, which was larger when  
 9 using the TOP-80 and RF feature sets (TOP-80 and RF = 94%) vs. TOP-60 (81%). An example  
 10 confusion matrix is shown Figure 9; in this case, the model had poorer performance in  
 11 classifying broad and narrow hand configurations compared to the one-hand configuration.

12

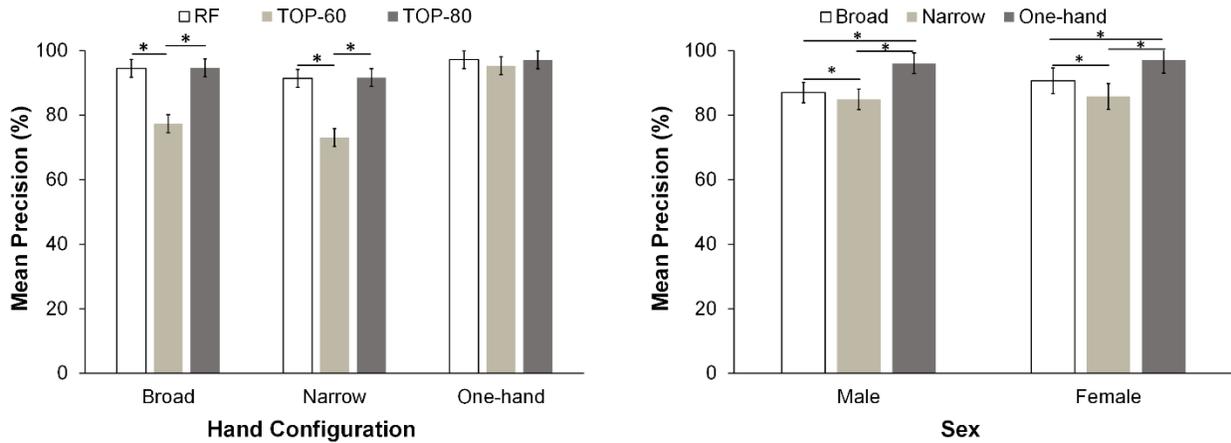


13

14 Figure 9: Overall confusion matrix with a Bi-LSTM model using the TOP-80 feature set,  
 15 representing the best performance when classifying *Hand Configuration*.

16 Precision: There were significant *RNN model*, *Feature set*, *Hand Configuration* main effects, as  
 17 well as *Feature set* × *Hand Configuration*, and *Hand Configuration* × *Sex* interaction effects on  
 18 precision (Table A.10). Using the GRU model led to significantly lower precision (89%),  
 19 compared to using the BGRU (90%) and Bi-LSTM models (91%), yet the magnitude of these  
 20 differences was small. Across hand configurations, using the TOP-60 feature sets led to  
 21 significantly less precision (by ~18%), compared to using the TOP-80 and RF feature sets,  
 22 except for the one-hand configuration (Figure 10). All simple effects were significant except for  
 23 the effects of *Feature set* in the one-hand ( $p = 0.089$ ) configuration. For a given sex, the narrow  
 24 hand configuration led to significantly lower precision (by up to 11%), compared to either the  
 25 broad or one-hand configurations (Figure 10). Simple effect analysis showed that precision  
 26 differed significantly across hand configurations within each sex ( $p < 0.0001$  for both males and

1 females), denoting important variations in precision depending on the specific hand  
 2 configuration.



3  
 4 Figure 10: Significant interaction effects of *Feature set* × *Hand Configuration* on precision  
 5 (Left) and of *Hand Configuration* × *Sex* on precision (Right).

6 Recall: There were significant *RNN model*, *Feature set*, *Hand Configuration* main effects and  
 7 *Feature set* × *Hand Configuration*, *Hand Configuration* × *Sex*, and *Feature set* × *Sex* × *Hand*  
 8 *Configuration* interaction effects on recall (Table A.10). Compared to the GRU model (88.9%),  
 9 using the Bi-LSTM led to significantly better recall (90.1%). Using the TOP-60 feature set led to  
 10 significantly lower recall in all three hand configurations (by 7-20%), compared to TOP-80 and  
 11 RF (Figure A.11). Of note, the magnitude of such differences differed between males and  
 12 females. All simple effects were significant, with differences between *Hand Configuration*  
 13 significant for both males ( $p < 0.0001$ ) and females ( $p < 0.0001$ ), and differences related to *Sex*  
 14 were significant for all *Hand Configurations* ( $p < 0.0001$ ).

15 F1-score: *RNN model*, *Feature set*, and *Hand Configuration* main effects, and *Feature set* ×  
 16 *Hand Configuration*, and *Feature set* × *Sex* interaction effects, were each significant (Table  
 17 A.10). Compared to the GRU model, using the Bi-LSTM model led to significantly higher F1-  
 18 scores, by up to 2%. Similar to results for recall, using the TOP-80 feature set led to significantly  
 19 higher F1-scores (93-97%), across all hand configurations, vs. RF and TOP-60 feature sets  
 20 (Figure A.12). All simple effects were significant, with differences between *Hand Configuration*  
 21 significant for all *Feature sets* ( $p = 0.0002$ ), and differences between *Feature set* were significant  
 22 for all *Hand Configuration* ( $p = 0.0015$ ). For a given *Sex*, using the TOP-80 feature set led to  
 23 significantly higher F1-scores (by ~12-15%) compared to TOP-60 (Figure 11). Simple effects  
 24 were significant for differences between *Hand Configuration* for both males ( $p < 0.0001$ ) and  
 25 females ( $p < 0.0001$ ).

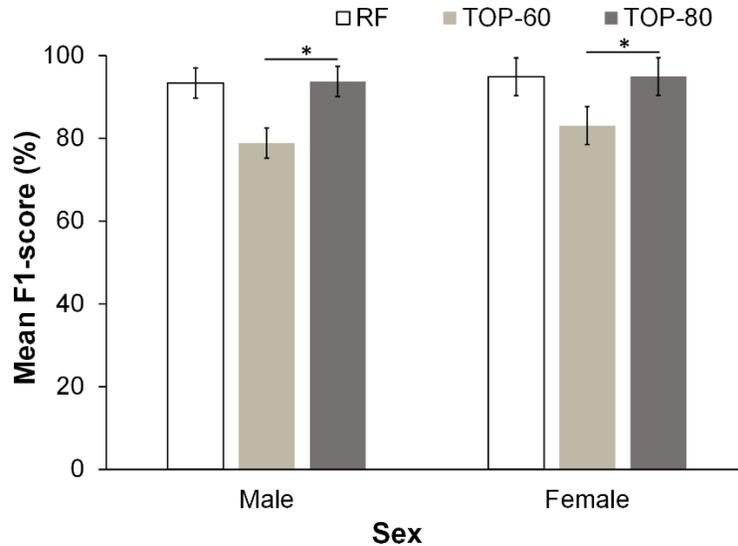


Figure 11: Significant interaction effect of *Feature set* × *Sex* on F1-score

### 3.3 Classifying Lift Origin

**Accuracy:** There were significant *RNN model* and *Feature set* main effects (Table A.8). Using the GRU model led to significantly poorer accuracy (~80%) vs. using Bi-LSTM (~83%) and BGRU models (~84%). Using the TOP-60 feature set led to significantly smaller accuracy (81%), compared to using RF (83%). Figure 12 displays an example confusion matrix for classifying lift origin.

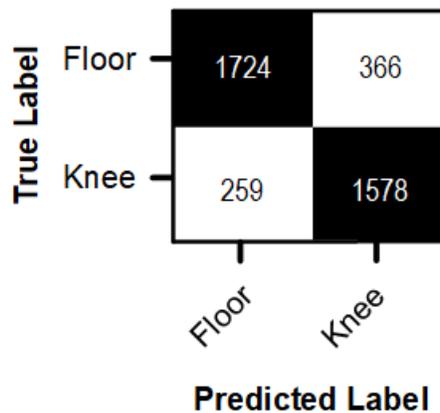
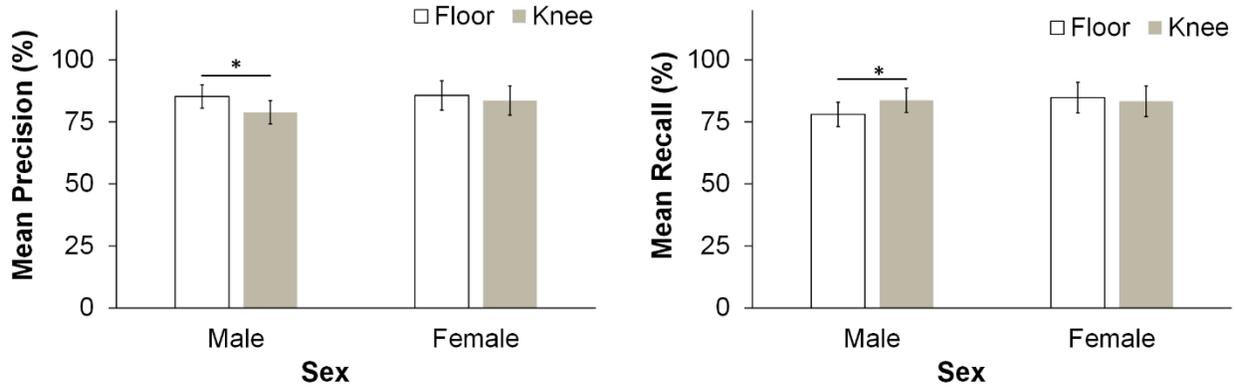


Figure 12: Overall confusion matrix with a BGRU model using the RF feature set, representing the best performance when classifying *Lift Origin*.

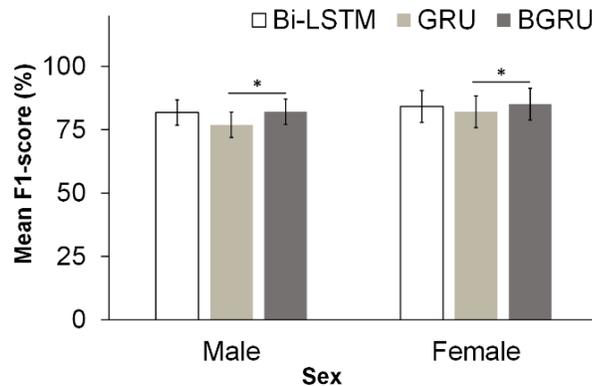
**Precision and Recall:** There were significant main and interaction effects of *RNN model*, *Lift Origin*, and *Sex* × *Lift Origin* (Table A.11). Using the GRU model led to significantly poorer precision (81%) vs. using both Bi-LSTM (84%) and BGRU (85%). Similarly, using the GRU model led to significantly lower recall (80%) vs. using the Bi-LSTM (83%) and BGRU (84%). Precision was lower among males (up to 5%) when lifting from the knee origin compared to the floor origin (Figure 13). Simple effects were significant for differences between *Lift Origin* for both males ( $p < 0.0001$ ) and females ( $p = 0.039$ ). Recall was also significantly lower among

1 males (by up to 5%) when performing lifts from the floor origin compared to the knee origin  
 2 (Figure 13). Simple effects were significant for differences between *Lift Origin* for both males ( $p$   
 3  $< 0.0001$ ) and females ( $p < 0.0001$ ).



4  
 5 Figure 13: Significant  $Sex \times Lift Origin$  interaction effects on precision (left) and recall (right)  
 6 for classifying *Lift Origin*.

7 F1-score: *RNN model*, *Feature set*, *Lift Origin* main effects and  $Sex \times RNN model$  interaction  
 8 effect were significant (Table A.11). Using RF and TOP-80 feature sets led to significantly larger  
 9 F1-scores (83% and 82%), compared to TOP-60 (80%). Lifting from the knee origin led to  
 10 significantly poorer F1-score (81%) compared to lifting from the floor origin (83%). Using the  
 11 BGRU model led to significantly larger F1-score compared to the GRU model, with an increase  
 12 of 5% for males and 3% for females (Figure 14). Simple effects were significant for the effect of  
 13 *Lift Origin* among males ( $p < 0.0001$ ) and females ( $p = 0.0032$ ).



14  
 15 Figure 14. Significant  $Sex \times RNN model$  interaction effect on F1-score for classifying *Lift Origin*.

#### 16 4.0 Discussion

17 Using data from an MMC system, our goal was to investigate the use of different RNN models  
 18 and feature sets in classifying diverse MMH tasks and specific task conditions. Across the MMH  
 19 task types and feature sets, mean precision, recall, and F1-score values were, in our opinion,  
 20 good to excellent, with each metric on the order of 85 – 97%. Performance in classifying hand

1 configuration was quite high, with mean precision, recall, and F1-scores of up to 96 – 97%.  
2 However, performance in classifying lift origin varied depending on the feature sets used and  
3 between males and females. The following discussion addresses these effects in more detail,  
4 including the differential effects of the three models and feature sets, performance dependencies  
5 on MMH task and task conditions, and differences related to sex.

#### 6 **4.1 Effects of Recurrent Neural Network Algorithms and Feature sets on MMH Task** 7 **Classification**

8 MMH task classification performance varied depending on the MMH task and feature set.  
9 Compared to the TOP-80 feature set, using the TOP-60 feature set substantially reduced mean  
10 precision and recall, by up to 7% depending on the specific MMH task. Recall that all  
11 participants were right-handed. Upon inspection (Tables A.1 – A.6), the TOP-60 features  
12 included tri-axial positions and quaternion joint rotations of the left forearm and shoulder, and  
13 the bilateral thigh, shin, and foot. TOP-80 features comprised tri-axial positions and quaternion  
14 joint rotations of the bilateral forearm and shoulder, along with tri-axial information on the left  
15 clavicle and quaternion joint rotations of the right clavicle. Given these differences, arm  
16 dominance might explain why the TOP-80 feature set outperformed the TOP-60 (aside from  
17 simply having additional input data). An individual’s dominant arm plays a crucial role in  
18 determining trajectory direction and speed, while the nondominant arm is critical for accurate  
19 positioning (Wang & Sainburg, 2007). Thus, including right-arm kinematics may have improved  
20 classification performance of deep learning algorithms.

21 Across the simulated MMH tasks, mean precision in task classification varied between 87 and  
22 97%, but depended on the specific RNN model and MMH task (Figure 8). Performance of deep  
23 learning models is often found to depend on the specific model and task type (Barazandeh et al.,  
24 2017; Luo et al., 2018; Porta et al., 2021). In such studies, a range of construction and MMH  
25 tasks were simulated, including steel bending, transporting, and lifting, and reported that deep  
26 learning algorithms often misclassify MMH tasks, especially when two tasks share comparable  
27 body kinematics (e.g., transporting vs. walking, lifting from knee vs. floor levels). We similarly  
28 found that deep learning algorithms misclassified tasks with comparable kinematics. Tasks  
29 involving pulling (Task 5) and carrying (Task 3) were the most “confused” tasks, irrespective of  
30 the deep learning algorithm (Figure A.5). Notably, most of the current misclassifications  
31 occurred at the completion of Task 3 and Task 5 (Figure 6). After reviewing all labeled videos,  
32 we suspect that the primary reason for such misclassification is the similarity in body kinematics  
33 between these tasks, especially at the end of these two tasks.

34 Another potential reason for misclassification is the presence of class imbalance particularly in  
35 Task 5. Class imbalance is a common issue in deep learning, where the instances of one class  
36 outnumber the instances of other classes (Guo et al., 2008). In some cases, this imbalance can  
37 lead to skewed performance, where precision and recall for a minority class are negatively  
38 affected, causing the model to misclassify minority class instances more frequently (Ali et al.,  
39 2013). In our dataset, Task 5 had the fewest sequences ( $N=2,779$ ), while Task 3 had the most  
40 sequences ( $N=6,391$ ). Thus, our deep learning algorithms might have biased towards predicting  
41 the majority class. Generating synthetic data using heuristic oversampling to increase Task 5  
42 occurrences could improve future classification performance. More specifically, heuristic  
43 oversampling techniques, such as the synthetic minority over-sampling technique (Chawla et al.,

1 2002), could be a promising approach to mitigate the effects of class imbalance, and improve the  
2 performance of deep learning algorithms in MMH task classification.

### 3 **4.2 Classifying Detailed Aspects of MMH Tasks**

4 Precision of hand configuration classifiers was substantially lower for the broad and narrow hand  
5 configurations (broad = 88.8%; narrow = 85.4%;), compared to the one-hand configuration  
6 (96.5%). This was an expected outcome, as similar motion patterns in the broad and narrow hand  
7 configurations could lead to redundant or correlated features, potentially affecting the ability of  
8 the models to generalize to new information during evaluation (Kim & Nussbaum, 2014).

9 Compared to Bi-LSTM and BGRU models, the GRU model showed substantially lower  
10 performance in classifying both hand configuration and lift origin, with mean precision decreases  
11 of up to 3% and 4%, respectively. The design of GRU models prioritizes computational  
12 efficiency, featuring simplified gates that process information only in the forward direction (i.e.,  
13 from past to future). However, this design may limit its effectiveness in situations where a  
14 comprehensive understanding of the entire sequence is crucial for making accurate  
15 classifications (Alawneh et al., 2020). The superior performance of Bi-LSTM and BGRU models  
16 is likely attributed to their capacity to process contextual information in bilateral temporal  
17 directions, thereby enabling a more comprehensive understanding of temporal sequences.

### 18 **4.3 Differences in Classifying MMH Task and Hand Configuration with Biological Sex**

19 There were differences related to sex in our study. For example, precision in some MMH tasks  
20 was 4% lower among males, albeit not a statistically significant difference. Also, mean recall was  
21 generally lower (up to 5%) when classifying lift origin among males. Frankly, it is unclear why  
22 our deep learning algorithms exhibited this sex-related bias, in particular, with the slightly lower  
23 classification performance among males. Ideally, a responsible deep learning algorithm should  
24 not be biased towards one or another group of people (Arrieta et al., 2020). deep learning  
25 algorithms, though, can harbor hidden biases that emerge when they are used in the real world.  
26 These so-called *latent biases* often inherited from the data used to train and validate deep  
27 learning algorithms, and which can pose a critical challenge to fairness and equity. For example,  
28 though not a direct comparison, sex-bias in training data led to a difference in emotion  
29 recognition accuracy between male and female test sets using RGB camera data (Domnich &  
30 Anbarjafari, 2021). A potential reason for sex-related differences in our results may stem from  
31 larger kinematic variability among males that could have affected the deep learning models'  
32 ability to accurately classify lift origin and some of the MMH tasks. Earlier studies have shown  
33 sex-related difference in kinematics when performing MMH tasks. For example, Lindbeck and  
34 Kjellberg (2001) documented that males exhibited larger kinematic variability (up to 20° trunk  
35 and knee flexion) when performing symmetric lifting from floor height. Plamondon et al. (2014)  
36 observed that females used sequential inter-joint coordination motion while males showed more  
37 synchronous motion when performing a repetitive palletizing task. Visual inspection of our video  
38 data also showed that males and females employed different postures, especially when  
39 performing Tasks 1, 5, and 8. We observed that females more often used a squat lifting  
40 technique, especially when completing symmetric box lifting (Task 1). Males in contrast, often  
41 adopted a mix of squat and stoop lifting. This difference may account for the smaller kinematic  
42 variability we found among the female participants. Thus, there may need a consideration of sex-

1 based biomechanical differences during MMH tasks in the development and validation of deep  
2 learning models.

### 3 **4.4 Limitations**

4 Several limitations of our study should be mentioned. Participants were relatively young (i.e., 18  
5 – 39 years old) and healthy. Therefore, caution should be taken in generalizing the results to  
6 other populations, such as older individuals or those with musculoskeletal disorders, and future  
7 work is needed with larger and more diverse samples. Feature selection methods can  
8 substantially affect classification performance (Preece et al., 2009). We used a filter-based  
9 approach to select subsets of features from among raw kinematic features. While our approach  
10 led to high performance in classifying MMH tasks and distinguishing among task conditions, the  
11 method used here – Minimal-Redundancy-Maximal-Relevancy – did not consider the temporal  
12 (time series) nature of the data. Future work should consider using feature selection methods that  
13 preserve temporal information. We used deep learning algorithms that consist of non-linear  
14 structures, making them highly non-transparent in arriving at their decisions. Understanding the  
15 process or reasoning behind predictions is crucial – a concept known as *Explainable AI*.  
16 Integrating Explainable AI into deep learning can facilitate verification of predictions, systematic  
17 identification of potential flaws and biases, and the understanding of the underlying decision-  
18 making processes of deep learning algorithms (Samek et al., 2017). We standardized some  
19 experimental aspects – such as participant clothing, lighting source and brightness, and camera  
20 positioning – which might have improved the ability to track whole-body kinematics using  
21 MMC. Thus, future work should explore the effects of varying these conditions.

22 Although our tasks included several common workplace elements (e.g., lifting height, box  
23 dimension), these tasks were performed in a fairly controlled laboratory setting. Some actual  
24 work environments often involve greater task variability and complexity with less controlled  
25 conditions. Therefore, caution should be exercised when generalizing our results to other tasks or  
26 work settings. Common components often seen in the workplace could lead to environmental  
27 occlusions, and three MMC systems were used to monitor whole-body kinematics to reduce the  
28 impact of such occlusions. Kotsifaki et al. (2018) found that increasing the number of cameras  
29 could enhance tracking of whole-body kinematics. Nevertheless, there will likely be practical  
30 constraints on the number of cameras that is feasible in practice, such as due to workflow and  
31 space restrictions. Therefore, future work should explore the possibility of using a single MMC  
32 or alternate configurations of two MMCs for physical exposure assessment. Recently, though,  
33 video surveillance cameras (plain cameras) are becoming more widely used in occupational  
34 settings (e.g., manufacturing) to enhance worker safety and to track productivity (Cocca et al.,  
35 2016; Kostal et al., 2022; Xu et al., 2015). We recommend investigating the feasibility of using  
36 plain cameras as an alternative approach to tracking body kinematics, which could be more  
37 efficient and less costly than optical or IMU systems (although the current MMC system was  
38 relatively inexpensive).

39 We removed data corresponding to idle time, since this part of the data included the calibration  
40 process and carrying the box from Task 5 to the cart, neither of which were among the MMH  
41 tasks of interest here. However, in practice, workers may perform various motions during idle  
42 times, which could be misinterpreted as MMH tasks. Future enhancements of our classification  
43 algorithms could include the ability to automatically filter out idle times and to differentiate

1 random, non-MMH motions from actual MMH tasks. Doing so would help to ensure more robust  
2 performance, even using untrimmed, real-world video data.

### 3 **4.5 Practical Applications of Our Findings**

4 We suggest that the GRU model and TOP-80 feature set could be used as the base approach for  
5 practical MMH task classification using MMC. Although the GRU model generally exhibited  
6 less accuracy (by 1%) compared to the Bi-LSTM and BGRU, this marginal difference seems of  
7 limited practical relevance. Further, training any of the RNN models with the TOP-80 feature set  
8 yielded a mean accuracy comparable to that using the RF features and roughly 2% better than  
9 when using the TOP-60 feature. Notably, the GRU model with the TOP-80 feature set required  
10 19 – 35% less mean epoch training time (the number of times that the model works through the  
11 entire training dataset) than when using the other models. This faster training time likely  
12 stemmed from the fact that the GRU model has a simplified architecture with two gates and  
13 fewer parameters, making it more computationally efficient for large datasets (Khandelwal et al.,  
14 2016). Moreover, using a streamlined feature set (vs. the RF feature set) helps promote a more  
15 interpretable model and contributes to improved generalization performance, by reducing the  
16 likelihood of capturing noise or irrelevant patterns in the data (Chen et al., 2017; Xue et al.,  
17 2015). Practitioners could also benefit from using a streamlined feature set, as continuous  
18 physical exposure assessments remain computationally expensive when using larger feature sets.  
19 Additionally, a streamlined feature set could inform the performance of our model when MMC is  
20 affected by occlusion and the RF feature set cannot be used.

21 Using RNN-based models and MMC is a novel approach to quantifying physical exposure  
22 assessment; as such, some comments about using this approach are warranted. One advantage is  
23 that using MMC provides a non-intrusive method for obtaining whole-body kinematics, and  
24 classification performance appears comparable to earlier results reported using inertial  
25 measurement units (IMUs). For example, earlier work that used Bi-LSTM and IMUs to classify  
26 MMH tasks reported a mean precision of 92% (Porta et al., 2021), while we achieved a mean  
27 precision of 91 – 92%. Moreover, one limitation of earlier work is that pushing and pulling tasks  
28 were often misclassified when using wearable sensors, by up to 5% (Kim & Nussbaum, 2014;  
29 Mokhlespour Esfahani, 2018; Porta et al., 2021). It is important to distinguish between pulling  
30 and pushing, though, since these tasks impose different loads on the low back (Hoozemans et al.,  
31 2004). Our approach was able to distinguish between pulling and pushing tasks with ambient  
32 sensing (i.e., MMC), with reasonable precision and recall (Figure 6). Using a sequence-to-  
33 sequence classification approach might have enhanced performance of our model, by capturing  
34 temporal dependencies and enabling RNN models to incorporate contextual information from the  
35 entire input sequence, in contrast to a sample-by-sample method (Yin et al., 2017).

### 36 **5.0 Conclusions**

37 Effectively measuring and monitoring physical exposures in the workplace is critical to assessing  
38 and controlling the risk of WMSDs. Several tools are available but are often limited in accuracy  
39 and scope, and their use can be quite resource-intensive. We evaluated the use of ambient sensors  
40 here, specifically MMC, together with RNNs, to classify among eight diverse MMH tasks and  
41 several specific task conditions. The classification of MMH tasks, using the MMC data,  
42 demonstrated rather robust performance across several classification models and input feature  
43 sets: mean precision, recall, and F1-score were 85 – 97%. Performance for classifying hand  
44 configuration was quite high, with mean precision, recall, and F1-scores of up to 97%. However,

1 performance in classifying lift origin varied substantially depending on feature sets and between  
2 males and females. Overall, our findings indicate that the proposed approach has the potential to  
3 efficiently and effectively quantify MMH tasks exposures, offering a balance of simplicity and  
4 non-intrusiveness in exposure assessments. Future work will be needed, though, to assess the  
5 ability of the proposed method among diverse workers and in real working conditions.

## 6 **6.0 Acknowledgements**

7 Support for this study was provided by the National Safety Council (NSC). The contents of this  
8 paper are solely the responsibility of the authors and do not necessarily represent the official  
9 views of NSC. We thank Ms. Sarah Iridiastadi for her assistance in data analysis.

## 1 7.0 REFERENCES

- 2 Alawneh, L., Mohsen, B., Al-Zinati, M., Shatnawi, A., & Al-Ayyoub, M. (2020). A comparison of  
3 unidirectional and bidirectional lstm networks for human activity recognition. In *2020 IEEE*  
4 *International Conference on Pervasive Computing and Communications Workshops (PerCom*  
5 *Workshops)* (pp. 1-6). IEEE.
- 6 Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J.*  
7 *Advance Soft Compu. Appl*, 5(3), 176-204.
- 8 Andersen, J. H., Kaergaard, A., Mikkelsen, S., Jensen, U. F., Frost, P., Bonde, J. P., Fallentin, N., & Thomsen,  
9 J. F. (2003). Risk factors in the onset of neck/shoulder pain in a prospective study of workers in  
10 industrial and service companies. *Occup Environ Med*, 60(9), 649-654.  
11 <https://doi.org/10.1136/oem.60.9.649>
- 12 Arisoy, E., Sethy, A., Ramabhadran, B., & Chen, S. (2015). Bidirectional recurrent neural network language  
13 models for automatic speech recognition. In *2015 IEEE International Conference on Acoustics,*  
14 *Speech and Signal Processing (ICASSP)* (pp. 5421-5425). IEEE.
- 15 Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S.,  
16 Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts,  
17 taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.
- 18 Barazandeh, B., Bastani, K., Rafieisakhaei, M., Kim, S., Kong, Z., & Nussbaum, M. A. (2017). Robust sparse  
19 representation-based classification using online sensor data for monitoring manual material  
20 handling tasks. *IEEE Transactions on Automation Science and Engineering*, 15(4), 1573-1584.
- 21 Bernard, B. P., & Putz-Anderson, V. (1997). Musculoskeletal disorders and workplace factors; a critical  
22 review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper  
23 extremity, and low back.
- 24 Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn  
25 inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3), 1-33.
- 26 Capela, N. A., Lemaire, E. D., & Baddour, N. (2015). Feature selection for wearable smartphone-based  
27 human activity recognition with able bodied, elderly, and stroke patients. *PLoS one*, 10(4),  
28 e0124414.
- 29 Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-  
30 sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
- 31 Chen, Q., Zhang, M., & Xue, B. (2017). Feature selection to improve generalization of genetic  
32 programming for high-dimensional symbolic regression. *IEEE Transactions on Evolutionary*  
33 *Computation*, 21(5), 792-806.
- 34 Chollet, F., & Others. (2015). Keras. <https://keras.io>
- 35 Ciriello, V. M., Snook, S. H., Hashemi, L., & Cotnam, J. (1999). Distributions of manual materials handling  
36 task parameters. *International Journal of Industrial Ergonomics*, 24(4), 379-388.  
37 [https://doi.org/Doi 10.1016/S0169-8141\(99\)00005-0](https://doi.org/Doi 10.1016/S0169-8141(99)00005-0)
- 38 Cocca, P., Marciano, F., & Alberti, M. (2016). Video surveillance systems to enhance occupational safety:  
39 A case study. *Safety Science*, 84, 140-148.
- 40 da Costa, B. R., & Vieira, E. R. (2010). Risk factors for work-related musculoskeletal disorders: A  
41 systematic review of recent longitudinal studies. *Am J Ind Med*, 53(3), 285-323.  
42 <https://doi.org/10.1002/ajim.20750>
- 43 David, G. C. (2005). Ergonomic methods for assessing exposure to risk factors for work-related  
44 musculoskeletal disorders. *Occup Med (Lond)*, 55(3), 190-199.  
45 <https://doi.org/10.1093/occmed/kqi082>
- 46 Dempsey, P. G., McGorry, R. W., & Maynard, W. S. (2005). A survey of tools and methods used by  
47 certified professional ergonomists. *Applied ergonomics*, 36(4), 489-503.

- 1 Djordjević, K. L., Jordović-Pavlović, M. I., Čojbašić, Ž., Galović, S., Popović, M. N., Nešić, M. V., &  
2 Markushev, D. D. (2022). Influence of data scaling and normalization on overall neural network  
3 performances in photoacoustics. *Optical and Quantum Electronics*, 54(8), 501.
- 4 Domnich, A., & Anbarjafari, G. (2021). Responsible AI: Gender bias assessment in emotion recognition.  
5 *arXiv preprint arXiv:2103.11436*.
- 6 Dwarampudi, M., & Reddy, N. (2019). Effects of padding on LSTMs and CNNs. *arXiv preprint*  
7 *arXiv:1903.07288*.
- 8 Escorcía, V., Dávila, M. A., Golparvar-Fard, M., & Niebles, J. C. (2012). Automated vision-based  
9 recognition of construction worker actions for building interior construction operations using  
10 RGBD cameras. Construction Research Congress 2012: Construction Challenges in a Flat World,
- 11 Garg, A., Chaffin, D. B., & Freivalds, A. (1982). Biomechanical stresses from manual load lifting: a static vs  
12 dynamic evaluation. *IIE transactions*, 14(4), 272-281.
- 13 Gary, A. W., Marras, W. S., & PARNIANPouR, M. (1996). Trunk kinematics of one-handed lifting, and the  
14 effects of asymmetry and load weight. *Ergonomics*, 39(2), 322-334.
- 15 Ghezlbash, F., Eskandari, A. H., Robert-Lachaine, X., Cao, S., Pesteie, M., Qiao, Z., Shirazi-Adl, A., &  
16 Larivière, C. (2024). Machine learning applications in spine biomechanics. *Journal of*  
17 *Biomechanics*, 111967.
- 18 Graves, A., Jaitly, N., & Mohamed, A. (2013). Hybrid speech recognition with Deep Bidirectional LSTM. In  
19 *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 273-278).  
20 <https://doi.org/10.1109/ASRU.2013.6707742>
- 21 Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. In *2008 Fourth*  
22 *international conference on natural computation* (Vol. 4, pp. 192-201). IEEE.
- 23 Han, S., Achar, M., Lee, S., & Peña-Mora, F. (2013). *Empirical assessment of a RGB-D sensor on motion*  
24 *capture and action recognition for construction worker monitoring*.  
25 <http://www.viejournal.com/content/1/1/6>
- 26 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*, 9(8), 1735-1780.  
27 <https://doi.org/10.1162/neco.1997.9.8.1735>
- 28 Hoozemans, M. J., Kuijjer, P. P. F., Kingma, I., Van Dieën, J. H., De Vries, W. H., Van Der Woude, L. H.,  
29 Veeger, D. J., Van Der Beek, A. J., & Frings-Dresen, M. H. (2004). Mechanical loading of the low  
30 back and shoulders during pushing and pulling activities. *Ergonomics*, 47(1), 1-18.
- 31 Jacobs, J. V., Hettinger, L. J., Huang, Y.-H., Jeffries, S., Lesch, M. F., Simmons, L. A., Verma, S. K., & Willetts,  
32 J. L. (2019). Employee acceptance of wearable technology in the workplace. *Applied ergonomics*,  
33 78, 148-156.
- 34 Jordao, A., Nazare Jr, A. C., Sena, J., & Schwartz, W. R. (2018). Human activity recognition based on  
35 wearable sensor data: A standardization of the state-of-the-art. *arXiv preprint arXiv:1806.05226*.
- 36 Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network  
37 architectures. International conference on machine learning,
- 38 Kent, P., Laird, R., & Haines, T. (2015). The effect of changing movement and posture using motion-sensor  
39 biofeedback, versus guidelines-based care, on the clinical outcomes of people with sub-acute or  
40 chronic low back pain-a multicentre, cluster-randomised, placebo-controlled, pilot trial. *BMC*  
41 *Musculoskeletal Disorders*, 16, 1-19.
- 42 Khandelwal, S., Lecouteux, B., & Besacier, L. (2016). *Comparing GRU and LSTM for automatic speech*  
43 *recognition*.
- 44 Khosrowpour, A., Niebles, J. C., & Golparvar-Fard, M. (2014). Vision-based workplace assessment using  
45 depth images for activity analysis of interior construction operations. *Automation in*  
46 *Construction*, 48, 74-87. <https://doi.org/10.1016/j.autcon.2014.08.003>

- 1 Kim, S., & Nussbaum, M. A. (2014). An evaluation of classification algorithms for manual material  
2 handling tasks based on data obtained using wearable technologies. *Ergonomics*, *57*(7), 1040-  
3 1051.
- 4 Kostal, P., Prajova, V., Vaclav, S., & Stan, S.-D. (2022). An Overview of the Practical Use of the CCTV System  
5 in a Simple Assembly in a Flexible Manufacturing System. *Applied System Innovation*, *5*(3), 52.
- 6 Kotsifaki, A., Whiteley, R., & Hansen, C. (2018). Dual Kinect v2 system can capture lower limb kinematics  
7 reasonably well in a clinical setting: concurrent validity of a dual camera markerless motion  
8 capture system in professional football players. *BMJ Open Sport Exerc Med*, *4*(1), e000441.  
9 <https://doi.org/10.1136/bmjsem-2018-000441>
- 10 le Feber, M., Jadoenathmisier, T., Goede, H., Kuijpers, E., & Pronk, A. (2021). Ethics and privacy  
11 considerations before deploying sensor technologies for exposure assessment in the workplace:  
12 results of a structured discussion amongst Dutch stakeholders. *Annals of work exposures and  
13 health*, *65*(1), 3-10.
- 14 LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (2002). Efficient backprop. In *Neural networks: Tricks of the  
15 trade* (pp. 9-50). Springer.
- 16 Levin, E. (1990). A recurrent neural network: Limitations and training. *Neural Networks*, *3*(6), 641-650.
- 17 Li, G., & Buckle, P. (1999). Current techniques for assessing physical exposure to work-related  
18 musculoskeletal risks, with emphasis on posture-based methods. *Ergonomics*, *42*(5), 674-695.  
19 <https://doi.org/10.1080/001401399185388>
- 20 Liberty Mutual Insurance. (2023). *2023 Workplace Safety Index: The Top 10 Causes of Disabling Injuries*.  
21 <https://business.libertymutual.com/insights/2023-workplace-safety-index/>
- 22 Lim, S., & D'Souza, C. (2020). A narrative review on contemporary and emerging uses of inertial sensing  
23 in occupational ergonomics. *International Journal of Industrial Ergonomics*, *76*, 102937.
- 24 Lind, C. M., Abtahi, F., & Forsman, M. (2023). Wearable motion capture devices for the prevention of  
25 work-related musculoskeletal disorders in ergonomics—an overview of current applications,  
26 challenges, and future opportunities. *Sensors*, *23*(9), 4259.
- 27 Lindbeck, L., & Kjellberg, K. (2001). Gender differences in lifting technique. *Ergonomics*, *44*(2), 202-214.
- 28 Logar, A. M., Corwin, E. M., & Oldham, W. J. (1993). A comparison of recurrent neural network learning  
29 algorithms. In *IEEE International Conference on Neural Networks* (pp. 1129-1134). IEEE.
- 30 Luo, H., Xiong, C., Fang, W., Love, P. E., Zhang, B., & Ouyang, X. (2018). Convolutional neural networks:  
31 Computer vision-based workforce activity assessment in construction. *Automation in  
32 Construction*, *94*, 282-289.
- 33 Lynn, H. M., Pan, S. B., & Kim, P. (2019). A deep bidirectional GRU network model for biometric  
34 electrocardiogram classification based on recurrent neural networks. *IEEE Access*, *7*, 145395-  
35 145405.
- 36 Markovitch, S., & Rosenstein, D. (2002). Feature generation using general constructor functions. *Machine  
37 Learning*, *49*, 59-98.
- 38 Marras, W. S., Cutlip, R. G., Burt, S. E., & Waters, T. R. (2009). National occupational research agenda  
39 (NORA) future directions in occupational musculoskeletal disorder health research. *Applied  
40 ergonomics*, *40*(1), 15-22.
- 41 Marras, W. S., & Davis, K. G. (1998). Spine loading during asymmetric lifting using one versus two hands.  
42 *Ergonomics*, *41*(6), 817-834.
- 43 Martinez, R., Bouffard, J., Michaud, B., Plamondon, A., Côté, J. N., & Begon, M. (2019). Sex differences in  
44 upper limb 3D joint contributions during a lifting task. *Ergonomics*, *62*(5), 682-693.
- 45 MassirisFernández, M., Fernández, J. Á., Bajo, J. M., & Delrieux, C. A. (2020). Ergonomic risk assessment  
46 based on computer vision and machine learning. *Computers & Industrial Engineering*, *149*,  
47 106816.

- 1 McNamara, R. J., Tsai, L. L. Y., Wootton, S. L., Ng, L. C., Dale, M. T., McKeough, Z. J., & Alison, J. A. (2016).  
2 Measurement of daily physical activity using the SenseWear Armband: Compliance, comfort,  
3 adverse side effects and usability. *Chronic respiratory disease*, *13*(2), 144-154.
- 4 Mokhlespour Esfahani, M. I. (2018). *Development and Assessment of Smart Textile Systems for Human*  
5 *Activity Classification* Virginia Tech].
- 6 Nath, N. D., Akhavian, R., & Behzadan, A. H. (2017). Ergonomic analysis of construction worker's body  
7 postures using wearable mobile sensors. *Appl Ergon*, *62*, 107-117.  
8 <https://doi.org/10.1016/j.apergo.2017.02.007>
- 9 Park, S., Park, J., Al-Masni, M. A., Al-Antari, M. A., Uddin, M. Z., & Kim, T.-S. (2016). A depth camera-  
10 based human activity recognition via deep learning recurrent neural network for health and  
11 social care services. *Procedia Computer Science*, *100*, 78-84.
- 12 Pedersen, S. J., Kitic, C. M., Bird, M.-L., Mainsbridge, C. P., & Cooley, P. D. (2016). Is self-reporting  
13 workplace activity worthwhile? Validity and reliability of occupational sitting and physical activity  
14 questionnaire in desk-based workers. *BMC Public Health*, *16*, 1-6.
- 15 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,  
16 Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine*  
17 *Learning research*, *12*, 2825-2830.
- 18 Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-  
19 dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and*  
20 *machine intelligence*, *27*(8), 1226-1238.
- 21 Plamondon, A., Larivière, C., Denis, D., Mecheri, H., Nastasia, I., & group, I. M. r. (2017). Difference  
22 between male and female workers lifting the same relative load when palletizing boxes. *Applied*  
23 *ergonomics*, *60*, 93-102.
- 24 Plamondon, A., Lariviere, C., Denis, D., St-Vincent, M., Delisle, A., & Group, I. M. R. (2014). Sex  
25 differences in lifting strategies during a repetitive palletizing task. *Applied Ergonomics*, *45*(6),  
26 1558-1569.
- 27 Plantard, P., Shum, H. P. H., Le Pierres, A. S., & Multon, F. (2017). Validation of an ergonomic assessment  
28 method using Kinect data in real workplace conditions. *Applied ergonomics*, *65*, 562-569.  
29 <https://doi.org/10.1016/j.apergo.2016.10.015>
- 30 Porta, M., Kim, S., Pau, M., & Nussbaum, M. A. (2021). Classifying diverse manual material handling tasks  
31 using a single wearable sensor. *Applied ergonomics*, *93*, 103386.
- 32 Preece, S. J., Goulermas, J. Y., Kenney, L. P., Howard, D., Meijer, K., & Crompton, R. (2009). Activity  
33 identification using body-mounted sensors—a review of classification techniques. *Physiological*  
34 *measurement*, *30*(4), R1.
- 35 Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random  
36 forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.
- 37 Punnett, L., & Wegman, D. H. (2004). Work-related musculoskeletal disorders: the epidemiologic  
38 evidence and the debate. *J Electromyogr Kinesiol*, *14*(1), 13-23.  
39 <https://doi.org/10.1016/j.jelekin.2003.09.015>
- 40 Roberts, D., Torres Calderon, W., Tang, S., & Golparvar-Fard, M. (2020). Vision-based construction worker  
41 activity analysis informed by body posture. *Journal of Computing in Civil Engineering*, *34*(4),  
42 04020017.
- 43 Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding,  
44 visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
- 45 Schall Jr, M. C., Sesek, R. F., & Cavuoto, L. A. (2018). Barriers to the adoption of wearable sensors in the  
46 workplace: A survey of occupational safety and health professionals. *Human Factors*, *60*(3), 351-  
47 362.

1 Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal*  
2 *Processing*, 45(11), 2673-2681.

3 Senator, T. E. (2005). Multi-stage classification. In *Fifth IEEE international conference on data mining*  
4 *(ICDM'05)* (pp. 8 pp.). IEEE.

5 Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification  
6 performance. *Applied Soft Computing*, 97, 105524.

7 Song, B., Kamal, A. T., Soto, C., Ding, C., Farrell, J. A., & Roy-Chowdhury, A. K. (2010). Tracking and activity  
8 recognition through consensus in distributed camera networks. *IEEE Trans Image Process*,  
9 19(10), 2564-2579. <https://doi.org/10.1109/TIP.2010.2052823>

10 Starbuck, R., Seo, J., Han, S., & Lee, S. (2014). A stereo vision-based approach to marker-less motion  
11 capture for on-site kinematic modeling of construction worker tasks. *Computing in Civil and*  
12 *Building Engineering - Proceedings of the 2014 International Conference on Computing in Civil*  
13 *and Building Engineering*,

14 Tang, S., & Golparvar-Fard, M. (2021). Machine Learning-Based Risk Analysis for Construction Worker  
15 Safety from Ubiquitous Site Photos and Videos. *Journal of Computing in Civil Engineering*, 35(6),  
16 04021020.

17 U.S. Bureau of Labor Statistics. (2021). *Nonfatal Occupational Injuries and Illnesses Requiring Days Away*  
18 *from Work*. bls.gov. Retrieved May 7 from <https://www.bls.gov/data/home.htm>

19 Wang, J., & Sainburg, R. L. (2007). The dominant and nondominant arms are specialized for stabilizing  
20 different features of task performance. *Experimental Brain Research*, 178, 565-570.

21 Wang, Y., Huang, J., He, T., & Tu, X. (2020). Dialogue intent classification with character-CNN-BGRU  
22 networks. *Multimedia Tools and Applications*, 79(7), 4553-4572.

23 Waters, T. R., Dick, R. B., Davis-Barkley, J., & Krieg, E. F. (2007). A cross-sectional study of risk factors for  
24 musculoskeletal symptoms in the workplace using data from the General Social Survey (GSS).  
25 *Journal of Occupational and Environmental Medicine*, 172-184.

26 Winter, D. A. (2009). Biomechanics and motor control of human movement. In (Third ed., pp. 49-50).  
27 John Wiley & Sons.

28 Xu, X., McGorry, R. W., Chou, L.-S., Lin, J.-h., & Chang, C.-c. (2015). Accuracy of the Microsoft Kinect™ for  
29 measuring gait parameters during treadmill walking. *Gait & posture*, 42(2), 145-151.

30 Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2015). A survey on evolutionary computation approaches to  
31 feature selection. *IEEE Transactions on Evolutionary Computation*, 20(4), 606-626.

32 Yang, J., Shi, Z., & Wu, Z. (2016). Vision-based action recognition of construction workers using dense  
33 trajectories. *Advanced Engineering Informatics*, 30(3), 327-336.

34 Yang, K., Ahn, C. R., & Kim, H. (2020). Deep learning-based classification of work-related physical load  
35 levels in construction. *Advanced Engineering Informatics*, 45, 101104-101104.  
36 <https://doi.org/10.1016/j.aei.2020.101104>

37 Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language  
38 processing. *arXiv preprint arXiv:1702.01923*.

39 Yu, Y., Yang, X., Li, H., Luo, X., Guo, H., & Fang, Q. (2019). Joint-Level Vision-Based Ergonomic Assessment  
40 Tool for Construction Workers. *Journal of Construction Engineering and Management*, 145(5),  
41 04019025. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001647](https://doi.org/10.1061/(asce)co.1943-7862.0001647)

42 Yun, S., Hong, S., Hwang, S., Lee, D., & Kim, H. (2025). Analysis of masonry work activity recognition  
43 accuracy using a spatiotemporal graph convolutional network across different camera angles.  
44 *Automation in Construction*, 175, 106178.

45 Zhan, K., Ramos, F., & Faux, S. (2012, 5-7 Dec. 2012). Activity recognition from a wearable camera. 2012  
46 12th International Conference on Control Automation Robotics & Vision (ICARCV),

47 Zhang, C., & Tian, Y. (2012). RGB-D camera-based daily living activity recognition. *Journal of Computer*  
48 *Vision and Image Processing*, 2(4), 12.

1 Zhang, H., Yan, X., & Li, H. (2018). Ergonomic posture recognition using 3D view-invariant features from  
2 single ordinary camera. *Automation in Construction, 94*, 1-10.  
3 <https://doi.org/10.1016/j.autcon.2018.05.033>  
4 Zhang, X., Schall Jr, M. C., Chen, H., Gallagher, S., Davis, G. A., & Sese, R. (2022). Manufacturing worker  
5 perceptions of using wearable inertial sensors for multiple work shifts. *Applied ergonomics, 98*,  
6 103579.  
7 Zhou, G., Aggarwal, V., Yin, M., & Yu, D. (2022). A Computer Vision Approach for Estimating Lifting Load  
8 Contributors to Injury Risk. *IEEE Transactions on Human-Machine Systems*.  
9 Zhou, X., Li, S., Liu, J., Wu, Z., & Chen, Y. F. (2024). Construction Activity Analysis of Workers Based on  
10 Human Posture Estimation Information. *Engineering, 33*, 225-236.

11

## Highlights

- Postures were continuously assessed using depth cameras
- Model performance varied in classifying manual material handling tasks and conditions
- Model performance was high when classifying hand configurations and lift origin
- Model performance varied depending on input feature sets
- Model performance varied across MMH tasks and differed between males and females
- The proposed approach has the potential to quantify MMH task exposures

1 **A Data-Driven Approach to Classifying Manual Material Handling Tasks Using Markerless**  
2 **Motion Capture and Recurrent Neural Networks**

3

4 <sup>a</sup>Aanuoluwapo Ojelade, <https://orcid.org/0000-0001-9715-3254>

5 <sup>b</sup>Mohammad Sadra Rajabi, <https://orcid.org/0000-0002-9100-3973>

6 <sup>b</sup>Sunwook Kim, <https://orcid.org/0000-0003-3624-1781>

7 <sup>b</sup>Maury A. Nussbaum, <https://orcid.org/0000-0002-1887-8431>

8

9 <sup>a</sup>Department of Industrial and Systems Engineering, University at Buffalo, Buffalo NY 14226,  
10 USA

11 <sup>b</sup>Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg VA 24061, USA

12

13

14 Corresponding address: Aanuoluwapo Ojelade

15 Department of Industrial and Systems Engineering,

16 University at Buffalo, 301 Bell Hall, Buffalo, NY 14226, USA.

17 Phone: 7166454721. Email: [aojelade@buffalo.edu](mailto:aojelade@buffalo.edu)

18

19

20 Acknowledgement: all authors have made substantial contributions to all of the following: (1)  
21 the conception and design of the study, or acquisition of data, or analysis and interpretation of  
22 data, (2) drafting the article or revising it critically for important intellectual content, (3) final  
23 approval of the version to be submitted.

24

25

1 **Abstract**

2 Work-related musculoskeletal disorders (WMSDs) are prevalent problems that encompass a  
3 range of conditions affecting muscles, tendons, and nerves due to repetitive strain, non-neutral  
4 postures, and forceful exertions. These disorders lead to pain, reduced productivity and  
5 substantial healthcare costs. Effective physical exposure assessment tools are needed in the  
6 workplace to quantify WMSD risks and the association between exposure and risks. While  
7 several tools are available, they are often limited in scope and lack the ability to assess physical  
8 risks continuously. In this study, we evaluated a data-driven approach to continuously classify  
9 manual material handling tasks and specific task conditions using different feature sets and  
10 machine learning algorithms. Specifically, kinematic data from markerless motion capture  
11 (MMC) system was used as input for various recurrent neural networks to classify among eight  
12 distinct manual material handling tasks: box lifting, asymmetric box lifting, box carriage, box  
13 pushing, box pulling, cart pushing, overhead lifting, and box lowering. The models we tested  
14 include bidirectional long-short term memory, gated recurrent units, and bidirectional gated  
15 recurrent units. We also classified specific task conditions, such as hand configurations and  
16 initial lifting height. Overall, using the MMC’s kinematic data led to satisfactory results (e.g.,  
17 accuracy of 80 – 94%) in classifying the tasks and the task conditions. Our results, though, also  
18 emphasize that classification performance varied across different feature sets, tasks, and between  
19 males and females. Nonetheless, use of MMC demonstrates clear potential for physical exposure  
20 assessment.

21

22 Keywords: Physical exposure assessment, Musculoskeletal disorders, Machine learning, sex  
23 differences, Computer vision

24

25

1 **1.0 Introduction**

2 Work-related musculoskeletal disorders (WMSDs) are injuries or dysfunctions affecting bones,  
3 nerves, tendons, muscles, and spinal discs (da Costa & Vieira, 2010). These injuries and  
4 degenerative conditions include nerve compression disorders, soreness, sprains, and strains  
5 (Punnett & Wegman, 2004). WMSDs continue to be an important occupational health concern.  
6 In the United States, WMSDs led to a median of 14 days away from work (U.S. Bureau of Labor  
7 Statistics, 2021) and involved substantial direct costs of about \$2.24 billion annually to  
8 employers (Liberty Mutual Insurance, 2023). Key risk factors contributing to the development of  
9 WMSDs include forceful exertions, repetition, and non-neutral postures. Exposures to such risk  
10 factors are particularly common during manual material handling (MMH) tasks, including  
11 lifting/lowering, pushing/pulling, and holding/carrying (Andersen et al., 2003; Bernard & Putz-  
12 Anderson, 1997; da Costa & Vieira, 2010). Effective assessment of physical exposures to such  
13 risks, though, is critical to developing targeted interventions, and more generally for quantifying  
14 associations between exposures and risks or doses (Marras et al., 2009; Plantard et al., 2017;  
15 Waters et al., 2007). A fundamental aspect of physical exposure assessment involves  
16 distinguishing the specific MMH tasks performed (i.e., task classification). This step is important  
17 in identifying the tasks associated with high-risk work conditions (Li & Buckle, 1999),  
18 especially since different tasks present distinct levels of physical exposures and associated  
19 WMSD risks.

20 Physical exposure assessments require methods that can classify MMH tasks accurately and that  
21 are compatible with the work environment. Physical exposure can be assessed using self-reports,  
22 human observations, and direct measurements (David, 2005; Li & Buckle, 1999). Self-report and  
23 human observation approaches are quick, straightforward, and require no advanced technologies.  
24 However, these approaches may be affected by individual biases or sub-optimal workplace  
25 conditions such as occlusions (Pedersen et al., 2016; Plantard et al., 2017). Alternatively, direct  
26 measurements often involve attaching sensors to objects or directly to a worker’s body, such as  
27 to obtain postural or force data (David, 2005; Lim & D'Souza, 2020). Direct measurements  
28 generally yield rich, precise data, but can be limited by cost, the time needed to analyze data,  
29 sensor discomfort, and potential influences on worker behaviors (Antwi-Afari et al., 2018;  
30 Golabchi et al., 2016; Nath et al., 2017). Overall, major limitations of direct measurement  
31 methods for continuous assessment are both labor intensive (Rezagholi et al., 2012) and resource  
32 demanding (Dianat et al., 2018; Wells et al., 1994). There is clear value in automating data  
33 collection and analysis to address the challenges posed by existing physical exposure assessment  
34 methods.

35 Recent advancements in sensor technology and machine learning have introduced new  
36 alternatives to perform occupational physical exposure assessments – specifically through  
37 computer-based methods (MassirisFernández et al., 2020). Of relevance here are *markerless*  
38 *motion capture* (MMC) systems, both plain and depth cameras. Existing work demonstrates the  
39 feasibility of a computer-based assessment approach for motion tracking and quantifying  
40 physical exposures, without requiring on-body sensors (Ghezelbash et al., 2024; Plantard et al.,  
41 2017; Roberts et al., 2020; Tang & Golparvar-Fard, 2021; Zhang et al., 2018; Zhou et al., 2024).  
42 Two example applications emphasize such feasibility. In one, data extracted from a MMC system  
43 were used to automatically detect activities such as walking and specific tasks during  
44 construction drywall installation (Khosrowpour et al., 2014). In another, MMC data facilitated

1 classifying different construction worker actions when laying bricks, transporting rebar, and  
2 making formwork (Yang et al., 2016).

3 However, reported applications of MMC have four important limitations. First, it is unclear if  
4 existing task classifiers can be applied when the tasks of interest include relatively complex  
5 motions. In both studies noted above (Khosrowpour et al., 2014; Yang et al., 2016), reasonable  
6 classification accuracy of ~75% was reported for construction tasks, such as shoveling and  
7 transporting, but accuracy substantially decreased for tasks that include similar body motions  
8 (e.g., bolting and plastering). Second, selecting an effective classification algorithm and input  
9 variables remains challenging, since classification performance can depend on the specific  
10 algorithm used and/or the inputs (aka *features*) used. Several task classification algorithms have  
11 been explored – such as Support Vector Machines, K-Nearest Neighbors, Decision Trees, and  
12 Neural Networks (e.g., Escorcía et al., 2012; Park et al., 2016; Song et al., 2010; Yu et al., 2019;  
13 Zhan et al., 2012; Zhang & Tian, 2012) – and performance for a given task clearly varies based  
14 on the specific algorithm and input variables employed. Third, the tasks included in earlier  
15 reports have often lack adequate representation of occupational tasks generally and MMH tasks  
16 specifically. Fourth, current computer-based assessments mainly classify discrete tasks, and it is  
17 unclear if they can accurately capture and analyze continuous or complex task sequences. This  
18 limitation is critical because discrete analysis can underestimate physical exposures in MMH  
19 tasks, highlighting the need for continuous quantification.

20 Our purpose in this study was thus to investigate the performance of an MMC system, together  
21 with machine learning algorithms, for classifying diverse MMH tasks during a simulated  
22 complex job. Specifically, we explored the relative performance of using different machine  
23 learning algorithms as an ergonomic exposure assessment approach for identifying specific  
24 MMH tasks and for distinguishing among different task conditions (e.g., initial lifting height).  
25 Several machine learning algorithms were tested, since no single model was expected to be best  
26 suited for MMH task classification (Jozefowicz et al., 2015). We sought to understand the  
27 performance of machine learning algorithms in capturing information in sequential MMH tasks.  
28 We selected RNN models over other classification algorithms (e.g., Support Vector Machines, K-  
29 Nearest Neighbors) given their superior performance in handling sequential data and in  
30 maintaining contextual relationships within segments of time series data (Arisoy et al., 2015;  
31 Logar et al., 1993; Schuster & Paliwal, 1997). Additionally, we evaluated the effects of various  
32 input variables (i.e., feature sets derived from kinematic data) on MMH task classification  
33 performance. Our study was exploratory in nature, seeking results that could inform future  
34 assessments of physical exposures using MMC. We expected the performance of machine  
35 learning algorithms to depend on the feature sets used, and to differ between specific MMH tasks  
36 and task conditions, and with biological sex. The latter expectation was based on evidence of  
37 kinematic differences in how males and females perform MMH tasks (Martinez et al., 2019;  
38 Plamondon et al., 2017).

## 39 **2.0 Methods**

40 MMH tasks were simulated in a controlled, laboratory setting, and these tasks were then  
41 classified using body kinematics obtained from an MMC system. Diverse tasks were simulated,  
42 representative of physically-demanding activities in several occupational sectors (e.g., lifting,  
43 carrying, pushing). Several task classifiers were explored using different machine learning

1 methods and feature sets, and the performance of these classifiers were examined using common  
2 metrics.

### 3 **2.1 Participants**

4 A convenience sample of 36 young (14 females) participants completed the study and were  
5 recruited from the university and local community. Respective means (SD) of age, body mass,  
6 and stature were 27 (4.4) years, 77.5 (12.2) kg, and 176.4 (6.9) cm for the males; and 27 (5.2)  
7 years, 68 (12.3) kg, and 170.1 (7.2) cm for the females. All participants self-reported being right-  
8 handed, physically active (i.e., exercising at least twice per week), and having no  
9 musculoskeletal disorders within the past year. The research reported herein complied with the  
10 tenets of the Declaration of Helsinki, and the study protocol was approved by the Institutional  
11 Review Board at Virginia Tech. Informed consent was obtained from all participants prior to any  
12 data collection.

### 13 **2.2 Task Simulations**

14 Eight MMH tasks were simulated in the laboratory, and these tasks involved some variations of  
15 manual box lifting, carrying, pushing, pulling, and reaching. The tasks simulated here are similar  
16 to those used in an earlier study that evaluated the efficacy of several classification models for  
17 MMH tasks (Kim and Nussbaum (2014)). Specifically, six within the current set of tasks were  
18 also included in the noted earlier study. Distinct here, however, was the inclusion of cart pushing  
19 and the use of different hand configurations, box masses, lift origins, and starting positions. The  
20 specific MMH tasks simulated are described below (see also Figures 1 and 3). A single wood box  
21 (width = 26.0 cm; depth = 41.0 cm; and height = 23.5 cm) was used across all tasks.

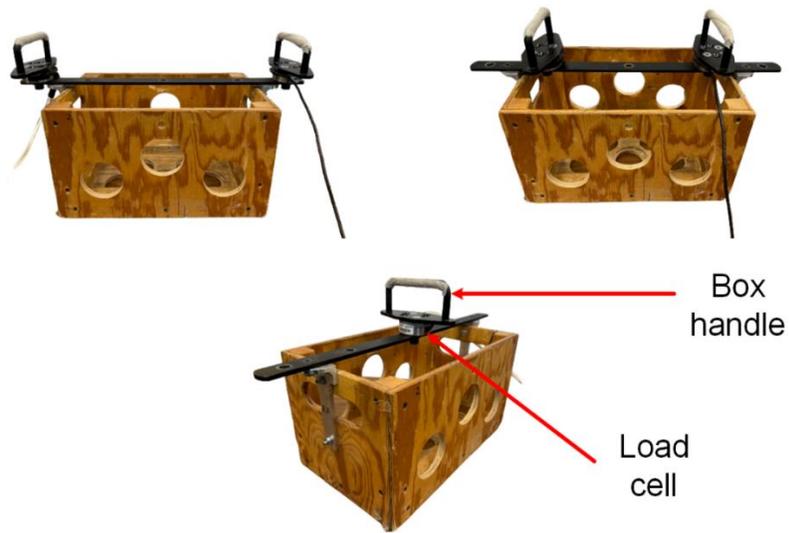
- 22 • Task 1: symmetric box lifting from two different origins (floor and individual knee  
23 height) to an individual hip height. Hip height was defined as the vertical distance from  
24 the floor to the greater trochanter.
- 25 • Task 2: asymmetric box lifting, from a table placed in front of the participant to another  
26 positioned 90° to the left of the participant. The two tables were adjusted to individual hip  
27 height.
- 28 • Task 3: box carriage from one table to another (i.e., lifting from hip height, carrying, and  
29 lowering to a height of 0.74m). Participants carried the box over a distance of 2.4 m,  
30 selected as the 50<sup>th</sup> percentile of the carrying distance of the U.S. workforce (Ciriello et al.,  
31 1999).
- 32 • Task 4: box pushing over a distance of ~0.7 m, with table height = 0.74 m.
- 33 • Task 5: box pulling toward the body over a distance of ~0.7m, with the table height fixed  
34 at 0.74 m.
- 35 • Task 6: cart pushing at individual waist height over a distance of ~1.8m. Cart mass was  
36 fixed at 86 kg, including the box, representing common cart loads (Hoozemans et al.,  
37 2004). Cart pushing was completed only using two hands.
- 38 • Task 7: overhead lifting from cart height to individual overhead height, with the box  
39 lifted from a height of 0.56 m (height of the cart). Overhead height was defined from  
40 individual anthropometric measures as the distance between the lateral epicondyle and  
41 the floor when the shoulder was flexed at 80°. This anthropometric measure was selected  
42 since working repeatedly with arm flexion or abduction beyond 80° has been associated  
43 with shoulder disorders (Bernard & Putz-Anderson, 1997).

- 1 • Task 8: lowering to different origins (floor and individual knee height) from overhead  
 2 height. Participants carried the box over a distance of ~1.3m.

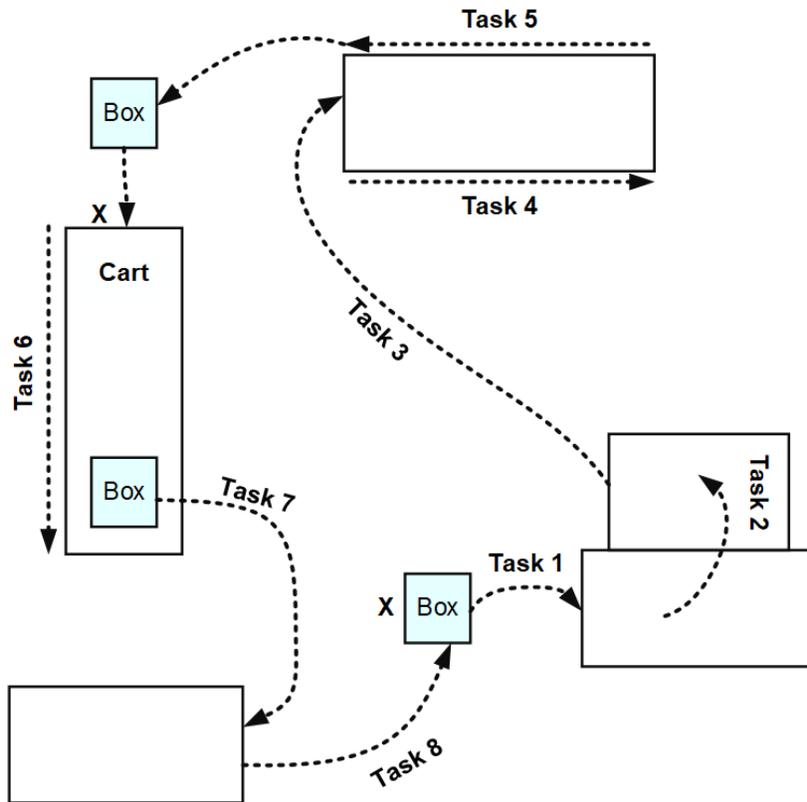


3  
 4 Figure 1: Illustrations of the simulated manual material handling tasks.

5 A total of 36 scenarios were completed, involving all possible combinations of the task  
 6 conditions: a) three levels of *Hand Configuration*; b) three levels of *Box Mass*; c) two levels of  
 7 *Lift Origin*; and d) two levels of *Start Position*. *Hand Configuration* had three levels: broad,  
 8 narrow, and one-hand (Figure 2). Both the broad and narrow hand configurations involved using  
 9 both hands, for which the handles were spaced at 52 and 33 cm apart, respectively. In the one-  
 10 hand configuration, the box handle was positioned in the middle of the box. Multiple hand  
 11 configurations were used here since different hand widths and one- vs. two-handed lifting  
 12 methods impose different biomechanical demands on the lower back (Garg et al., 1982; Gary et  
 13 al., 1996; Marras & Davis, 1998). Three levels of *Box Mass* – 6, 9, and 12 kg – were used for the  
 14 broad and narrow hand configuration, while masses of 5, 7, and 9 kg were used for one-hand  
 15 hand configuration. Different box masses for one vs. two hands lifting tasks were selected, based  
 16 on results from pilot testing (i.e., to ensure that most participants could complete all tasks).  
 17 Specific masses used here were roughly within the 8<sup>th</sup> and 22<sup>nd</sup> percentiles of masses lifted by the  
 18 U.S. workforce (Ciriello et al., 1999). Two levels of *Lift Origin* – floor and individual knee  
 19 height – were included to impose different physical exposures during lifting. Finally, two levels  
 20 of *Starting Position* were used to impose more task variability. Specifically, one starting position  
 21 was set at Task 1 and the other at Task 6 (Figure 3).



1  
 2 Figure 2: Illustration of the three levels of hand configuration: broad (top-left), narrow (top-  
 3 right), and one-hand (bottom-middle). Note that the handles were oriented parallel to the short  
 4 sides of the box (they appear at an angle in the figure only as an artifact of the lens setting used).



5  
 6 Figure 3: Top-view schematic of the simulated tasks. Dotted lines indicate the movement path,  
 7 and "X" indicates two alternative starting positions. Descriptions of the eight tasks are provided  
 8 in the text.

## 1    **2.3 Experimental Procedures**

2    Participants completed one experimental session (~3 hrs.), which consisted of training and  
3    experimental phases. During the training phase, participants were introduced to the MMH tasks  
4    and the different task conditions, then practiced the simulated tasks. They were asked to perform  
5    all tasks using their own comfortable work strategies and speed, while assuming they were  
6    working in an industrial environment.

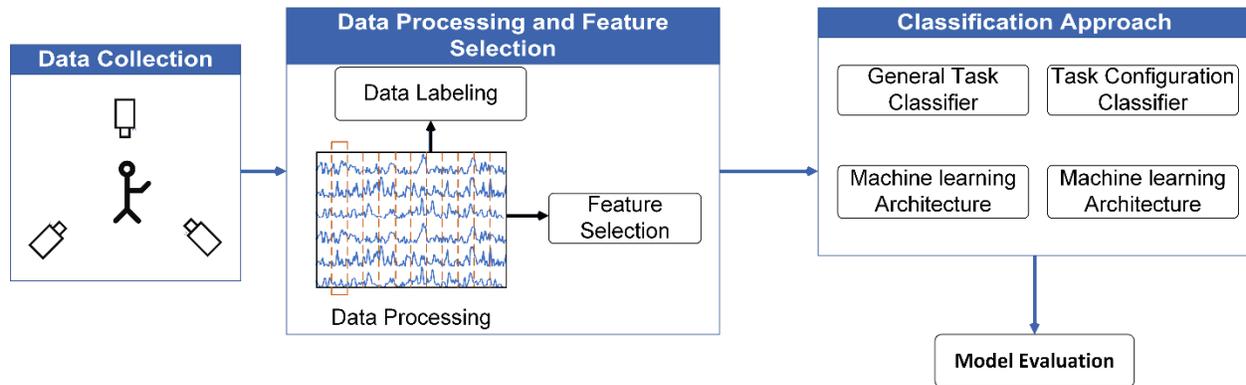
7    In the experimental phase, participants completed multiple trials of the MMH tasks, performing  
8    one trial for each of the 36 task conditions. A study *trial* involved completing all of the eight  
9    MMH tasks sequentially, with a given box mass, hand configuration, and lift origin, and from a  
10   given starting position. The presentation order of *Hand Configuration* was counterbalanced using  
11    $3 \times 3$  balanced Latin Squares. Within a given *Hand Configuration*, the presentation order of *Box*  
12   *Mass* was also counterbalanced using  $3 \times 3$  balanced Latin Squares, whereas the presentation  
13   orders of *Starting Position* and *Lift Origin* were alternated across participants. To mitigate  
14   physical fatigue, a minimum of four minutes of rest was given between each *Hand Position*  
15   condition.

## 16   **2.4 Instrumentation**

17   Whole-body kinematics were monitored at 30 Hz using three markerless camera systems (Azure  
18   Kinect™, Microsoft Corporation, Seattle WA, USA). Note that the Azure Kinect is the latest  
19   depth camera from Microsoft, and it has clear improvements such as a wider field of view, better  
20   resolution, and a global shutter that allows for improved performance in sunlight (Microsoft  
21   Corporation, 2022). These systems were positioned ~1.74 m from the edge of the work area, a  
22   configuration that was determined to be effective during pilot testing and that aimed to optimize  
23   the coverage of the narrow camera field of view (see Figure Appendix.1 or A.1). The three Azure  
24   Kinects were time-synchronized using a 3.5 mm auxiliary cable connected in a daisy-chain  
25   configuration, where one Azure Kinect was designated as the primary device, with the remaining  
26   two as secondary devices. An iPi Recorder (iPi Soft®; [www.ipisoft.com](http://www.ipisoft.com)) was used for sampling  
27   from the Azure Kinects. At the start of each trial, participants completed a calibration step by  
28   standing in a “T-pose” while facing one of the markerless cameras. This required them standing  
29   upright, with their arms abducted horizontally, and their feet together and pointing forward. This  
30   calibration was used to establish a clear and consistent starting point for tracking body joints.

## 31   **2.5 Data Processing and Feature Selection**

32   Task classification can be viewed as multi-staged processing following data collection: data  
33   processing, feature selection, data labeling, classification approach, and model evaluation (Figure  
34   4). Each of these stages is discussed in detail subsequently.



1

2

Figure 4: Overview of the task classification process.

3

### 2.5.1 Data Processing

4 Data recorded from the markerless camera systems were processed using iPi Motion Capture  
 5 Studio. Using this software, 3D body motions were tracked, and key body joints were identified  
 6 from the video recorded using the iPi Recorder. Seventeen body segments were tracked and  
 7 stored for further processing – pelvis; lower, middle, and upper spine; neck and bilateral  
 8 acromioclavicular; shoulders, elbows, hips, knees, and taluses. Tri-axial positions and quaternion  
 9 joint rotations were extracted for each of the body joints using the iPi Biomechanical add-on  
 10 software. Joint kinematics were low-pass filtered (6 Hz cutoff; 4<sup>th</sup> order Butterworth;  
 11 bidirectional) to remove sensor noise and other artifacts, with the cutoff frequency determined  
 12 using residual analysis (Winter, 2009). Filtered joint kinematics were then normalized, using the  
 13 Min-Max-Scaler function (Pedregosa et al., 2011), to a (0, 1) range across participants, given that  
 14 machine learning models are sensitive to unscaled data (Djordjević et al., 2022; LeCun et al.,  
 15 2002; Singh & Singh, 2020). All subsequent offline data processing was completed using Python  
 16 (ver. 3.10.11; <https://www.python.org>.)

17

### 2.5.2 Feature Selection

18 Features are independent measurable properties or characteristics of the data that serve as input  
 19 for machine learning models. In their basic form, features can be raw data. In some instances,  
 20 however, the most informative features are selected while discarding irrelevant or redundant  
 21 ones, by using feature selection algorithms (Markovitch & Rosenstein, 2002). Therefore, we  
 22 considered two types of features: raw and “informative” features. First, the processed joint  
 23 kinematics, which included 119 features from the processed kinematics (i.e., raw features = RF),  
 24 consisting of tri-axial positions and quaternion elements of the 17 body joints, were included as  
 25 input for model training described below. Second, high-dimensional features may lead to longer  
 26 classification processes and overfitting. Reducing dimensionality, by selecting the most  
 27 informative features, can simplify machine learning models and help improve model  
 28 performance and interpretability (Chen et al., 2017). Dimensionality reduction is especially  
 29 valuable in the context of MMC systems, since occlusions from the body or the environment can  
 30 lead to data loss and inaccurate pose estimations (Plantard et al., 2017). A filter-based method for  
 31 feature selection – Minimal-Redundancy-Maximum-Relevancy (Peng et al., 2005) – was used to  
 32 select subsets of features from among the entire kinematic feature pool or RF. The primary goal  
 33 of this approach is to find a feature subset that minimizes redundancy between features, while  
 34 maximizing their relevance to the target variable, and more details on this method have been

1 reported elsewhere (Peng et al., 2005). Since there is no consensus on the number of features that  
2 could yield the best model performance, we arbitrarily tested the top 60 (TOP-60) and top 80  
3 (TOP-80) features as input for each machine learning model. Listings of these two feature sets  
4 for each classification scheme (see Section 2.6.3) are provided in Tables A.1 – A.6.

### 5 **2.5.3 Data Labeling**

6 Processed data were labeled manually for the specific tasks performed by visually observing the  
7 recorded RGB-D data (served as the ground truth). In each MMH task that involved  
8 manipulating a box, the task began when participants touched the box handle and ended when  
9 they removed their hands from the box. In the case of cart pushing (Task 6), the task started  
10 when participants touched the cart handles and ended when they removed their hands from the  
11 box. Data corresponding to when participants were idle between each task were removed, and  
12 the remaining data were concatenated (see example in Figure A.3).

## 13 **2.6 Classification Approach**

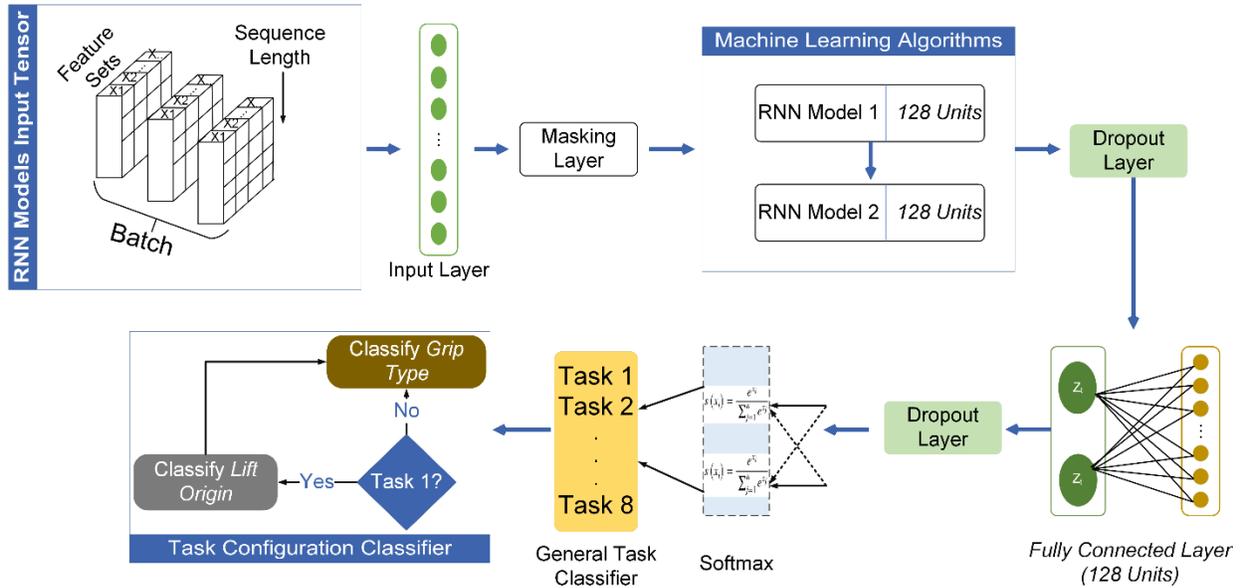
### 14 **2.6.1 Machine Learning Algorithms**

15 Recurrent neural network (RNN) architectures were used to classify MMH tasks using MMC-  
16 derived kinematic measures. RNNs, which are a variant of an artificial deep learning network,  
17 are designed to process and recognize patterns in sequential data (Levin, 1990). Unlike other  
18 neural networks and machine learning models that process data in a cross-sectional way, RNNs  
19 maintain contextual relationships over time series segments. Three RNN models were used and  
20 compared here for classifying the MMH tasks: 1) Bidirectional Long Short-Term Memory (Bi-  
21 LSTM); 2) Gated Recurrent Units (GRU); and 3) Bidirectional Gated Recurrent Units (BGRU).  
22 Bi-LSTM, which is a variant of LSTM, incorporates both forward and backward LSTM models  
23 to process data in both directions, effectively addressing the vanishing gradient problem  
24 (Hochreiter & Schmidhuber, 1997). This problem occurs when the gradients of the loss function  
25 become extremely small, causing the network weights to update very slowly (Hochreiter &  
26 Schmidhuber, 1997; Jozefowicz et al., 2015). The gating mechanisms of a Bi-LSTM includes  
27 three gates (i.e., input, forget, and output) and allows for modifications of Bi-LSTM cell states.  
28 By integrating inputs from past-to-future and future-to-past directions, this model enhances task  
29 classification performance (Graves et al., 2013; Yang et al., 2020; Zhou et al., 2022). A GRU  
30 model simplifies the LSTM model by using fewer parameters, improving its efficiency in  
31 *understanding* long-term dependencies. Finally, a BGRU is built upon the GRU model by adding  
32 a bi-directional layer, providing the model output layer with complete contextual information of  
33 the input data at each time point. In brief, the input data are passed through feedforward and  
34 backward GRU networks, and the outputs of these two pathways are connected at the same outer  
35 layer. Some studies have shown that BGRU models are suitable for classification problems, such  
36 as for human identification (Lynn et al., 2019) and for dialog intent classification (Wang et al.,  
37 2020).

38

### 39 **2.6.2 Model Architecture**

40 The model architecture consisted of an input layer, an RNN model layer (i.e., Bi-LSTM, GRU,  
41 or BGRU), a masking layer, dropout functions, dense layers, and an output layer (Figure 5).



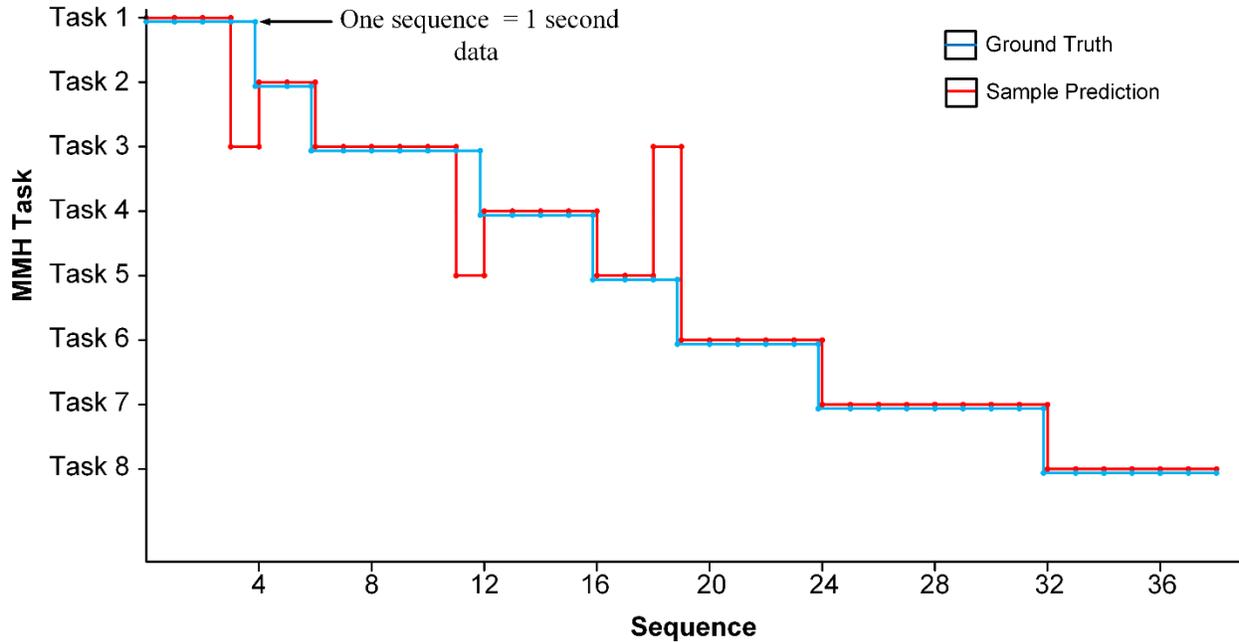
1

2

Figure 5: Overall architecture of recurrent neural network models for classification.

3 The input layer of each RNN model was designed to take a multidimensional matrix defined by  
 4 sequence length, feature dimension, and batch size. Sequence refers to the ordered set of input  
 5 data points processed sequentially over time steps, which was fixed at 30 here (Figure A.2),  
 6 corresponding to 1 second of data at the 30 Hz sampling rate. This sequence length was selected  
 7 to account for tasks that have short duration, and it has been used in previous human activity  
 8 recognition studies to improve recognition of short duration tasks (Bulling et al., 2014; Capela et  
 9 al., 2015). To account for the varying duration of each task, the sequence was zero-padded  
 10 (Figure A.2, Appendix A) to ensure a consistent length (e.g., see (Dwarampudi & Reddy, 2019)).  
 11 A masking layer was added to each of the RNN models to mitigate padding effects. Specifically,  
 12 this layer skips any sequences that have the special masking value. Feature dimension is the  
 13 number of features, based on the dimensions of the features generated earlier. Raw, TOP-60, and  
 14 TOP-80 features were 119-, 60- and 80-dimensional vectors, respectively. Finally, the batch size  
 15 was set to 64, which was determined as the optimal size during initial testing. The input layer  
 16 was connected to each of the RNN model architectures, consisting of a dropout function, a dense  
 17 layer, an optimizer, a loss function, and an output classification layer. Our output classification  
 18 layer used a many-to-one architecture, wherein a single output is synthesized from the input data.

19 Classification decisions were made *sequence-to-sequence*, specifically one decision for each  
 20 second of data. Each input sequence was classified independently, with the entire sequence being  
 21 considered at every time step (Figure 6). In this context, a time step corresponds to each discrete  
 22 unit of time within an input data sequence, and here the discrete time unit was 1 second. The  
 23 number of sequences varied across tasks due to differences in task completion time and the way  
 24 sequences were defined. Using time steps enables RNNs to capture temporal dependencies and  
 25 facilitates modeling of dynamic input sequences. Time steps also enable RNN models to  
 26 maintain contextual relationships between past and present information, making RNNs well-  
 27 suited for tasks involving time series data.



1  
 2 Figure 6: An example of continuous classification using a sequence-to-sequence approach. A Bi-  
 3 LSTM model and Raw feature set were used in this example. Each dot represents a sequence,  
 4 blue lines indicate ground truth, and red lines indicate output from the classification model. In  
 5 this example, three classification errors are evident.

### 6 **2.6.3 Classification Scheme**

7 Two categories of RNN models were developed – *general task classifiers* and *task configuration*  
 8 *classifiers*. General task classifiers were trained using each of the three feature sets (RF, TOP-60,  
 9 and TOP-80) to classify the eight simulated MMH tasks. Task configuration classifiers, in  
 10 contrast, classified the different configurations within relevant simulated tasks (i.e., Lift Origins  
 11 and/or Hand Configuration). We used a multi-stage classification approach using two  
 12 classification stages because using a single classifier for classifying both tasks and task  
 13 configurations could result in more frequent false positives (Senator, 2005).

### 14 **2.6.4 Model Training, Validation, Hyperparameters, and Evaluation**

15 For each category of RNN model, experimental data were used as input to the models. We used a  
 16 leave-one-subject-out approach to train and validate the RNN models; this approach is a special  
 17 case of cross-validation, in which each subject is considered as a “fold”. Thus, data from 35  
 18 participants were used for training, with the remaining participant’s data used for validation. This  
 19 process was repeated 36 times (i.e., 36-fold cross-validation). While the leave-one-subject-out  
 20 method has been used in past work (e.g., Kim & Nussbaum, 2014; Porta et al., 2021), it often  
 21 results in high variance in accuracy since participants can perform the same tasks in different  
 22 ways (Jordao et al., 2018). This method, however, replicates real-world training and testing,  
 23 wherein the model is trained offline using known subject data and then tested on an unseen  
 24 subject (Jordao et al., 2018).

25 RNN models have several hyperparameters that are used to control the learning process and  
 26 model complexity (Probst et al., 2019). Hyperparameter values were determined here using an

1 empirical tuning process involving manual adjustments to each parameter and subsequent  
2 evaluation of the resulting model performance. Specific values tested and the final values  
3 adopted are presented in Table A.7. Performance of the RNN models was assessed using four  
4 common metrics – macro accuracy (accuracy), Precision, recall, and F1-score – which are  
5 provided in Figure A.4. Macro accuracy was assessed to account for the class imbalance across  
6 MMH tasks.

## 7 **2.7 Statistical Analyses**

8 To account for the differing number of independent variables that could affect each performance  
9 metric, two sets of analyses of variance (ANOVAs) models were used. First, a two-way repeated-  
10 measures ANOVA was used to assess the effects of *Feature set* and *RNN model* on accuracy. The  
11 latter effect had three levels, to represent general task, hand configuration, and lift origin  
12 classifiers. This initial analysis was performed, since accuracy could only be computed across all  
13 eight simulated MMH tasks. Second, separate three-way repeated-measures ANOVAs were used  
14 for the remaining performance metrics (precision, recall, and F1-score). For general task  
15 classification, the independent variables were *Feature set*, *RNN model*, and *MMH task*. Two sets  
16 of models were used for the task configuration classifiers, with different independent variables:  
17 a) *Feature set*, *RNN models*, and *Hand Configuration*; and b) *Feature set*, *RNN models*, and *Lift*  
18 *Origin*. For each ANOVA model, biological sex (*Sex*) was included as a blocking effect,  
19 significant interaction effects were explored using simple-effects testing, and *post hoc* paired  
20 comparisons were completed using the Tukey’s HSD procedure. All statistical analyses were  
21 performed with JMP Pro 16 (SAS, Cary, NC) using the restricted maximum likelihood (REML)  
22 method. Parametric model assumptions were verified, and statistical significance was determined  
23 when  $p < 0.05$ . Summary data are reported as least-square means (with 95% confidence  
24 intervals) based on the statistical model fits.

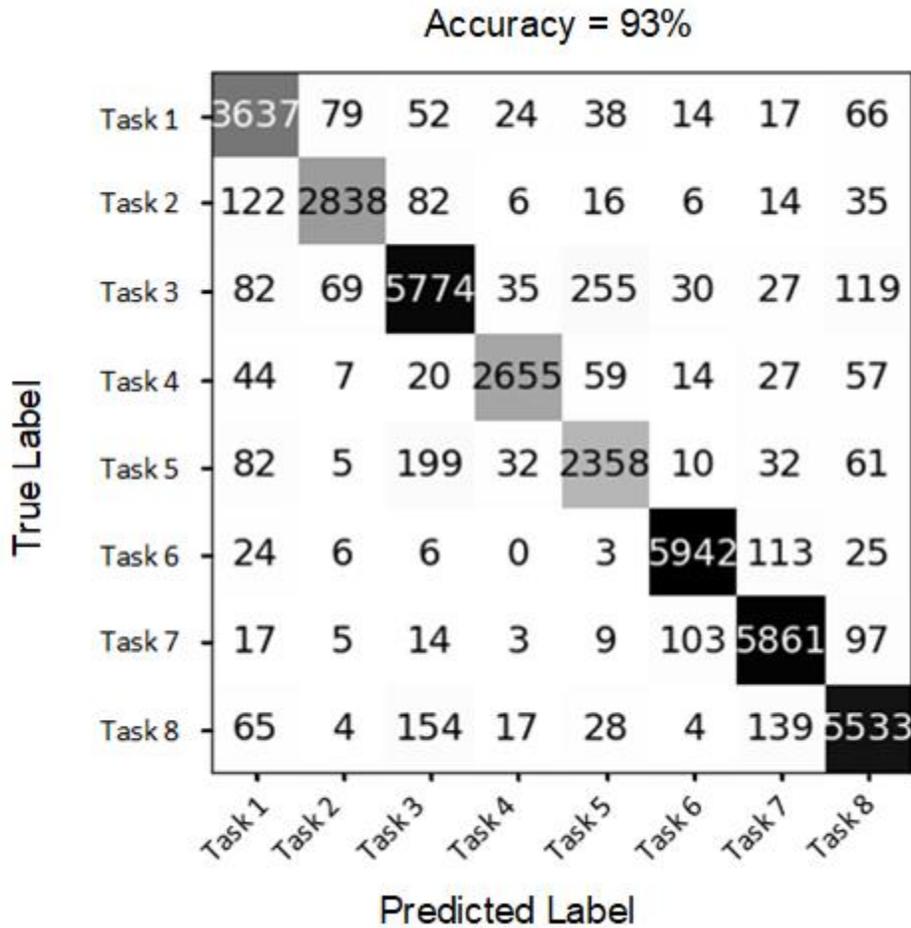
## 25 **3.0 Results**

26 ANOVA results are summarized in Tables A.8 – A.11, and Figures A.5 – A.7 provides confusion  
27 matrices for each category of RNN model. Sample confusion matrices are shown below (Figures  
28 6, 8 and 10) using results representing the best classification performance we obtained. There  
29 were significant main or interactive effects of *RNN model* and *Feature set* for all classification  
30 performance metrics. More detailed results are provided below.

### 31 **3.1 MMH Task Classification Performance**

32 Accuracy: There were significant main effects of both *RNN model* and *Feature set* on accuracy  
33 (Table A.8). The GRU model yielded significantly smaller accuracy (~91%), vs. the BGRU and  
34 Bi-LSTM models (~92%), though the magnitude of the difference was clearly quite small. Using  
35 the TOP-60 feature set led to significantly lower mean GRU accuracy compared to the TOP-80  
36 and RF feature sets, though again the difference was rather small (~90 vs. ~93%, respectively). A  
37 sample confusion matrix is shown in Figure 7; in this case, the model performed well on most  
38 tasks, with an overall accuracy of ~93%.

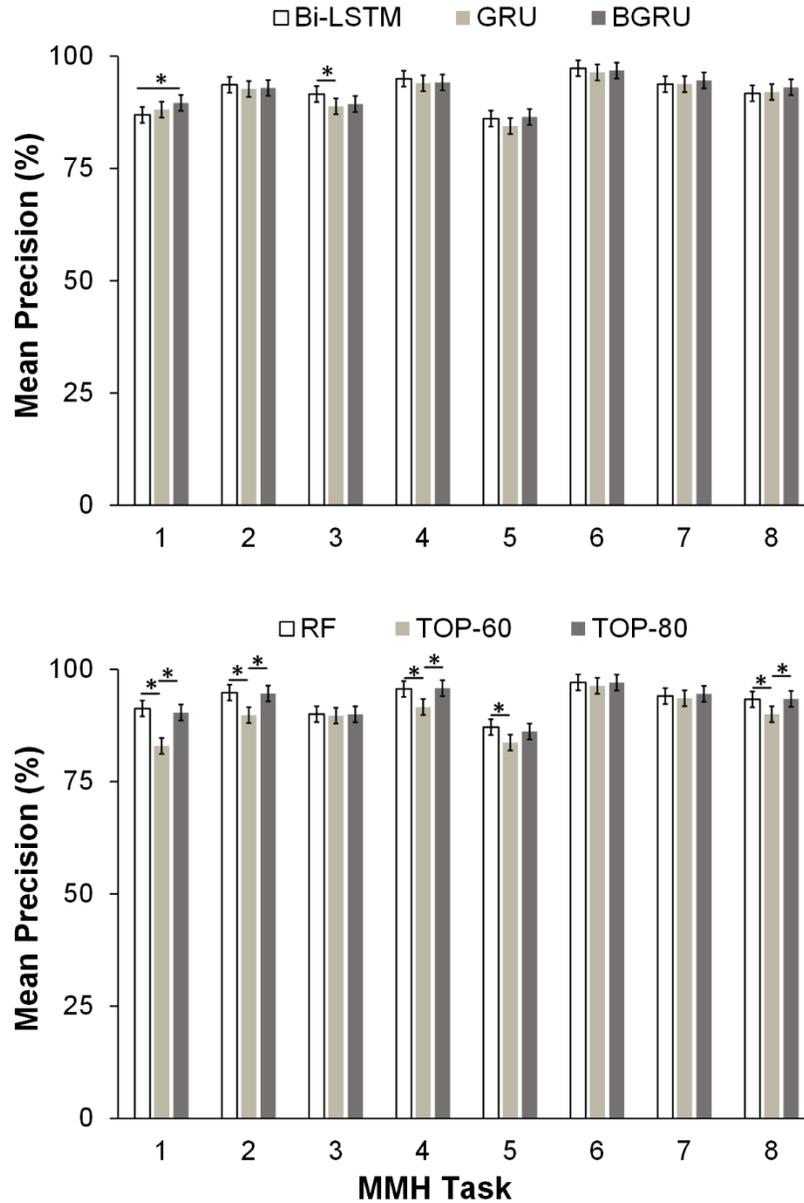
39



1

2 Figure 7: Overall confusion matrix from using a Bi-LSTM model using the TOP-80 feature set,  
 3 representing the best performance when classifying the MMH tasks. For this and other confusion  
 4 matrices, cells on the main diagonal indicate correct classifications, and lighter shades indicate  
 5 more misclassifications.

6 Precision: *RNN model*, *MMH task*, and *Feature set* main effects, and *RNN model* × *MMH task*,  
 7 *Feature set* × *MMH task*, *Sex* × *MMH task* interaction effects, were all significant (Table A.9).  
 8 Precision was relatively high and comparable between the different RNN models for most of the  
 9 MMH tasks (i.e., ~87-97%; Figure 8). However, using the GRU model led to precision that was  
 10 up to ~3% lower compared to the Bi-LSTM and BGRU models (Figure 8). Simple effects were  
 11 significant, except for the effect of *RNN model* in Tasks 2 ( $p = 0.37$ ), 4 ( $p = 0.29$ ), 6 ( $p = 0.37$ ),  
 12 and 7 ( $p = 0.36$ ). Using the TOP-60 feature set consistently resulted in 2-8% lower precision than  
 13 when using the TOP-80 and RF feature sets, with this difference depending on the specific MMH  
 14 task (Figure 8). Simple effects were significant, except for the effect of *Feature set* in Tasks 3 ( $p$   
 15 = 0.87), 6 ( $p = 0.46$ ), and 7 ( $p = 0.36$ ). Precision in some tasks was ~4% smaller among males  
 16 than females, though no significant paired differences between sexes were found for any tasks  
 17 (Figure A.8). Simple effect analysis showed that precision differed significantly across MMH  
 18 tasks within each sex ( $p < 0.0001$  for both males and females), indicating the specific MMH task  
 19 performed influenced precision for both sexes.



1

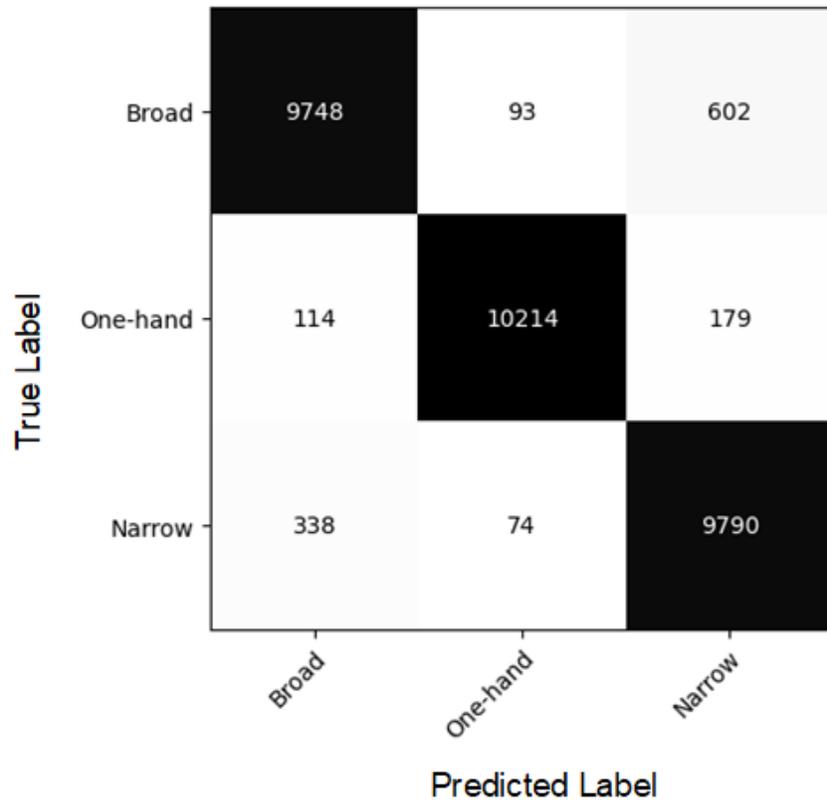
2 Figure 8: Interaction effects of *RNN Model*  $\times$  *MMH task* (top) and *Feature set*  $\times$  *MMH task*  
 3 (bottom) on classification *precision*. For this and other figures below, error bars indicate 95%  
 4 confidence intervals, and the symbol \* indicates a significant difference between pairs of means.

5 Recall and F1-score: RNN model, MMH task, and Feature set main effects, and Feature set  $\times$   
 6 MMH task and Sex  $\times$  MMH task interaction effects, were all significant for both recall and F1-  
 7 score (Table A.9; Figures A.9 and A.10). One consistent observation was that using the TOP-60  
 8 features led to significantly reduced recall and F1-scores compared to TOP-80 and RF, by up to  
 9 7% depending on the specific MMH task (Figures A.9 and A.10). All simple effects were  
 10 significant, except for the effect of Feature set in Tasks 6 ( $p = 0.51$ ) and 7 ( $p = 0.20$ ). No  
 11 significant paired differences were observed between sexes or between MMH tasks. As was the  
 12 case for precision, though, these metrics were somewhat lower among males, by  $\sim 1$ – $3\%$   
 13 depending on the specific MMH task. Simple effect analysis showed that recall and F1-score

1 differed significantly across MMH tasks within each sex ( $p < 0.0001$  for both males and  
2 females).

### 3 **3.2 Classifying Hand Configuration**

4 Accuracy: There was a significant main effect of *Feature set* on accuracy, which was larger when  
5 using the TOP-80 and RF feature sets (TOP-80 and RF = 94%) vs. TOP-60 (81%). An example  
6 confusion matrix is shown Figure 9; in this case, the model had poorer performance in  
7 classifying broad and narrow hand configurations compared to the one-hand configuration.

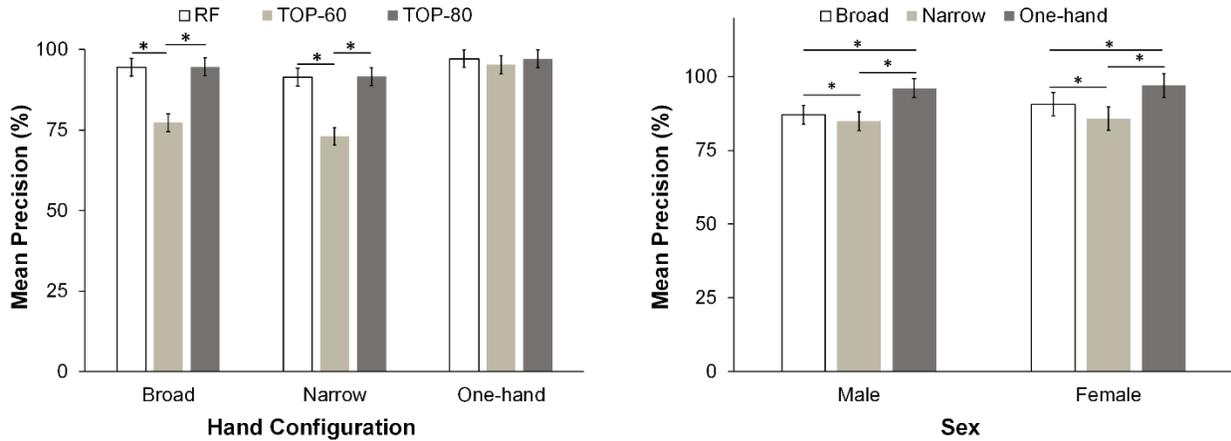


8

9 Figure 9: Overall confusion matrix with a Bi-LSTM model using the TOP-80 feature set,  
10 representing the best performance when classifying *Hand Configuration*.

11 Precision: There were significant *RNN model*, *Feature set*, *Hand Configuration* main effects, as  
12 well as *Feature set*  $\times$  *Hand Configuration*, and *Hand Configuration*  $\times$  *Sex* interaction effects on  
13 precision (Table A.10). Using the GRU model led to significantly lower precision (89%),  
14 compared to using the BGRU (90%) and Bi-LSTM models (91%), yet the magnitude of these  
15 differences was small. Across hand configurations, using the TOP-60 feature sets led to  
16 significantly less precision (by  $\sim 18\%$ ), compared to using the TOP-80 and RF feature sets,  
17 except for the one-hand configuration (Figure 10). All simple effects were significant except for  
18 the effects of *Feature set* in the one-hand ( $p = 0.089$ ) configuration. For a given sex, the narrow  
19 hand configuration led to significantly lower precision (by up to 11%), compared to either the  
20 broad or one-hand configurations (Figure 10). Simple effect analysis showed that precision  
21 differed significantly across hand configurations within each sex ( $p < 0.0001$  for both males and

1 females), denoting important variations in precision depending on the specific hand  
 2 configuration.



3  
 4 Figure 10: Significant interaction effects of *Feature set* × *Hand Configuration* on precision  
 5 (Left) and of *Hand Configuration* × *Sex* on precision (Right).

6 Recall: There were significant *RNN model*, *Feature set*, *Hand Configuration* main effects and  
 7 *Feature set* × *Hand Configuration*, *Hand Configuration* × *Sex*, and *Feature set* × *Sex* × *Hand*  
 8 *Configuration* interaction effects on recall (Table A.10). Compared to the GRU model (88.9%),  
 9 using the Bi-LSTM led to significantly better recall (90.1%). Using the TOP-60 feature set led to  
 10 significantly lower recall in all three hand configurations (by 7-20%), compared to TOP-80 and  
 11 RF (Figure A.11). Of note, the magnitude of such differences differed between males and  
 12 females. All simple effects were significant, with differences between *Hand Configuration*  
 13 significant for both males ( $p < 0.0001$ ) and females ( $p < 0.0001$ ), and differences related to *Sex*  
 14 were significant for all *Hand Configurations* ( $p < 0.0001$ ).

15 F1-score: *RNN model*, *Feature set*, and *Hand Configuration* main effects, and *Feature set* ×  
 16 *Hand Configuration*, and *Feature set* × *Sex* interaction effects, were each significant (Table  
 17 A.10). Compared to the GRU model, using the Bi-LSTM model led to significantly higher F1-  
 18 scores, by up to 2%. Similar to results for recall, using the TOP-80 feature set led to significantly  
 19 higher F1-scores (93-97%), across all hand configurations, vs. RF and TOP-60 feature sets  
 20 (Figure A.12). All simple effects were significant, with differences between *Hand Configuration*  
 21 significant for all *Feature sets* ( $p = 0.0002$ ), and differences between *Feature set* were significant  
 22 for all *Hand Configuration* ( $p = 0.0015$ ). For a given *Sex*, using the TOP-80 feature set led to  
 23 significantly higher F1-scores (by ~12-15%) compared to TOP-60 (Figure 11). Simple effects  
 24 were significant for differences between *Hand Configuration* for both males ( $p < 0.0001$ ) and  
 25 females ( $p < 0.0001$ ).

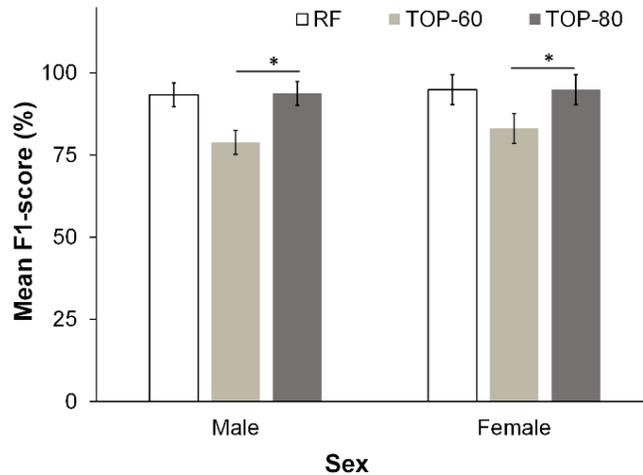


Figure 11: Significant interaction effect of *Feature set* × *Sex* on F1-score

### 3.3 Classifying Lift Origin

Accuracy: There were significant *RNN model* and *Feature set* main effects (Table A.8). Using the GRU model led to significantly poorer accuracy (~80%) vs. using Bi-LSTM (~83%) and BGRU models (~84%). Using the TOP-60 feature set led to significantly smaller accuracy (81%), compared to using RF (83%). Figure 12 displays an example confusion matrix for classifying lift origin.

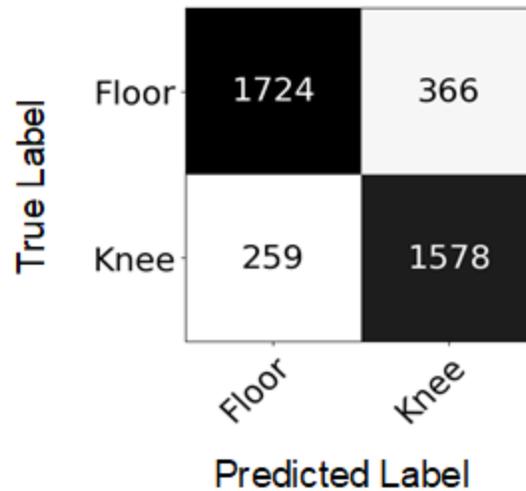
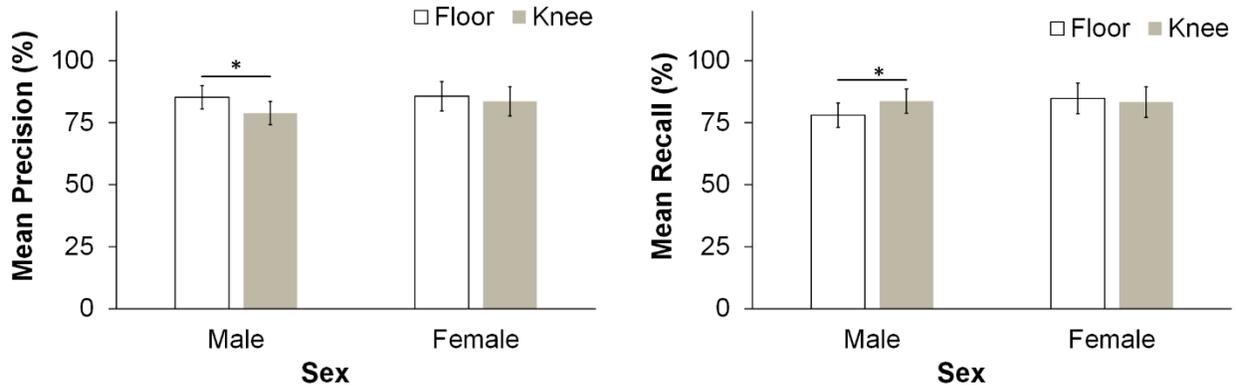


Figure 12: Overall confusion matrix with a BGRU model using the RF feature set, representing the best performance when classifying *Lift Origin*.

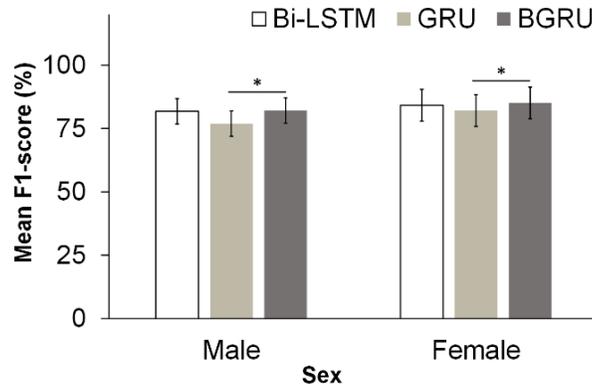
Precision and Recall: There were significant main and interaction effects of *RNN model*, *Lift Origin*, and *Sex* × *Lift Origin* (Table A.11). Using the GRU model led to significantly poorer precision (81%) vs. using both Bi-LSTM (84%) and BGRU (85%). Similarly, using the GRU model led to significantly lower recall (80%) vs. using the Bi-LSTM (83%) and BGRU (84%). Precision was lower among males (up to 5%) when lifting from the knee origin compared to the floor origin (Figure 13). Simple effects were significant for differences between *Lift Origin* for

1 both males ( $p < 0.0001$ ) and females ( $p = 0.039$ ). Recall was also significantly lower among  
 2 males (by up to 5%) when performing lifts from the floor origin compared to the knee origin  
 3 (Figure 13). Simple effects were significant for differences between *Lift Origin* for both males ( $p$   
 4  $< 0.0001$ ) and females ( $p < 0.0001$ ).



5  
 6 Figure 13: Significant  $Sex \times Lift Origin$  interaction effects on precision (left) and recall (right)  
 7 for classifying *Lift Origin*.

8 F1-score: RNN model, Feature set, Lift Origin main effects and  $Sex \times RNN model$  interaction  
 9 effect were significant (Table A.11). Using RF and TOP-80 feature sets led to significantly larger  
 10 F1-scores (83% and 82%), compared to TOP-60 (80%). Lifting from the knee origin led to  
 11 significantly poorer F1-score (81%) compared to lifting from the floor origin (83%). Using the  
 12 BGRU model led to significantly larger F1-score compared to the GRU model, with an increase  
 13 of 5% for males and 3% for females (Figure 14). Simple effects were significant for the effect of  
 14 *Lift Origin* among males ( $p < 0.0001$ ) and females ( $p = 0.0032$ ).



15  
 16 Figure 14. Significant  $Sex \times RNN model$  interaction effect on F1-score for classifying *Lift Origin*.

#### 17 4.0 Discussion

18 Using data from an MMC system, our goal was to investigate the use of different RNN models  
 19 and feature sets in classifying diverse MMH tasks and specific task conditions. Across the MMH  
 20 task types and feature sets, mean precision, recall, and F1-score values were, in our opinion,

1 good to excellent, with each metric on the order of 85 – 97%. Performance in classifying hand  
2 configuration was quite high, with mean precision, recall, and F1-scores of up to 96 – 97%.  
3 However, performance in classifying lift origin varied depending on the feature sets used and  
4 between males and females. The following discussion addresses these effects in more detail,  
5 including the differential effects of the three models and feature sets, performance dependencies  
6 on MMH task and task conditions, and differences related to sex.

#### 7 **4.1 Effects of Machine Learning Algorithms and Feature sets on MMH Task Classification**

8 MMH task classification performance varied depending on the MMH task and feature set.  
9 Compared to the TOP-80 feature set, using the TOP-60 feature set substantially reduced mean  
10 precision and recall, by up to 7% depending on the specific MMH task. Recall that all  
11 participants were right-handed. Upon inspection (Tables A.1 – A.6), the TOP-60 features  
12 included tri-axial positions and quaternion joint rotations of the left forearm and shoulder, and  
13 the bilateral thigh, shin, and foot. TOP-80 features comprised tri-axial positions and quaternion  
14 joint rotations of the bilateral forearm and shoulder, along with tri-axial information on the left  
15 clavicle and quaternion joint rotations of the right clavicle. Given these differences, arm  
16 dominance might explain why the TOP-80 feature set outperformed the TOP-60 (aside from  
17 simply having additional input data). An individual’s dominant arm plays a crucial role in  
18 determining trajectory direction and speed, while the nondominant arm is critical for accurate  
19 positioning (Wang & Sainburg, 2007). Thus, including right-arm kinematics may have improved  
20 classification performance of machine learning algorithms.

21 Across the simulated MMH tasks, mean precision in task classification varied between 87 and  
22 97%, but depended on the specific RNN model and MMH task (Figure 8). Performance of  
23 machine learning models is often found to depend on the specific model and task type  
24 (Barazandeh et al., 2017; Gong et al., 2011; Khosrowpour et al., 2014; Luo et al., 2018; Yang et  
25 al., 2016). In such studies, a range of construction and MMH tasks were simulated, including  
26 drilling, sawing, and lifting, and reported that machine learning algorithms often misclassify  
27 MMH tasks, especially when two tasks share comparable body kinematics (e.g., drilling vs.  
28 sawing, lifting from knee vs. floor levels). We similarly found that machine learning algorithms  
29 misclassified tasks with comparable kinematics. Tasks involving pulling (Task 5) and carrying  
30 (Task 3) were the most “confused” tasks, irrespective of the machine learning algorithm (Figure  
31 A.5). Notably, most of the current misclassifications occurred at the completion of Task 3 and  
32 Task 5 (Figure 6). After reviewing all labeled videos, we suspect that the primary reason for such  
33 misclassification is the similarity in body kinematics between these tasks, especially at the end of  
34 these two tasks.

35 Another potential reason for misclassification is the presence of class imbalance particularly in  
36 Task 5. Class imbalance is a common issue in machine learning, where the instances of one class  
37 outnumber the instances of other classes (Guo et al., 2008). In some cases, this imbalance can  
38 lead to skewed performance, where precision and recall for a minority class are negatively  
39 affected, causing the model to misclassify minority class instances more frequently (Ali et al.,  
40 2013). In our dataset, Task 5 had the fewest sequences ( $N=2,779$ ), while Task 3 had the most  
41 sequences ( $N=6,391$ ). Thus, our machine learning algorithms might have biased towards  
42 predicting the majority class. Generating synthetic data using heuristic oversampling to increase  
43 Task 5 occurrences could improve future classification performance. More specifically, heuristic  
44 oversampling techniques, such as the synthetic minority over-sampling technique (Chawla et al.,

1 2002), could be a promising approach to mitigate the effects of class imbalance, and improve the  
2 performance of machine learning algorithms in MMH task classification.

### 3 **4.2 Classifying Detailed Aspects of MMH Tasks**

4 Precision of hand configuration classifiers was substantially lower for the broad and narrow hand  
5 configurations (broad = 88.8%; narrow = 85.4%;), compared to the one-hand configuration  
6 (96.5%). This was an expected outcome, as similar motion patterns in the broad and narrow hand  
7 configurations could lead to redundant or correlated features, potentially affecting the ability of  
8 the models to generalize to new information during evaluation (Kim & Nussbaum, 2014).

9 Compared to Bi-LSTM and BGRU models, the GRU model showed substantially lower  
10 performance in classifying both hand configuration and lift origin, with mean precision decreases  
11 of up to 3% and 4%, respectively. The design of GRU models prioritizes computational  
12 efficiency, featuring simplified gates that process information only in the forward direction (i.e.,  
13 from past to future). However, this design may limit its effectiveness in situations where a  
14 comprehensive understanding of the entire sequence is crucial for making accurate  
15 classifications (Alawneh et al., 2020). The superior performance of Bi-LSTM and BGRU models  
16 is likely attributed to their capacity to process contextual information in bilateral temporal  
17 directions, thereby enabling a more comprehensive understanding of temporal sequences.

### 18 **4.3 Differences in Classifying MMH Task and Hand Configuration with Biological Sex**

19 There were differences related to sex in our study. For example, precision in some MMH tasks  
20 was 4% lower among males, albeit not a statistically significant difference. Also, mean recall was  
21 generally lower (up to 5%) when classifying lift origin among males. Frankly, it is unclear why  
22 our machine learning algorithms exhibited this sex-related bias, in particular, with the slightly  
23 lower classification performance among males. Ideally, a responsible machine learning algorithm  
24 should not be biased towards one or another group of people (Arrieta et al., 2020). Machine  
25 learning algorithms, though, can harbor hidden biases that emerge when they are used in the real  
26 world. These so-called *latent biases* often inherited from the data used to train and validate  
27 machine learning algorithms, and which can pose a critical challenge to fairness and equity. For  
28 example, though not a direct comparison, sex-bias in training data led to a difference in emotion  
29 recognition accuracy between male and female test sets using RGB camera data (Domnich &  
30 Anbarjafari, 2021). A potential reason for sex-related differences in our results may stem from  
31 larger kinematic variability among males that could have affected the machine learning models'  
32 ability to accurately classify lift origin and some of the MMH tasks. Earlier studies have shown  
33 sex-related difference in kinematics when performing MMH tasks. For example, Lindbeck and  
34 Kjellberg (2001) documented that males exhibited larger kinematic variability (up to 20° trunk  
35 and knee flexion) when performing symmetric lifting from floor height. Plamondon et al. (2014)  
36 observed that females used sequential inter-joint coordination motion while males showed more  
37 synchronous motion when performing a repetitive palletizing task. Visual inspection of our video  
38 data also showed that males and females employed different postures, especially when  
39 performing Tasks 1, 5, and 8. Thus, there may need a consideration of sex-based biomechanical  
40 differences during MMH tasks in the development and validation of machine learning models.

41

42

## 1 4.4 Limitations

2 Several limitations of our study should be mentioned. First, participants were relatively young  
3 (i.e., 18 – 39 years old) and healthy. Therefore, caution should be taken in generalizing the  
4 results to other populations, such as older individuals or those with musculoskeletal disorders,  
5 and future work is needed with larger and more diverse samples. Second, feature selection  
6 methods can substantially affect classification performance (Preece et al., 2009). We used a  
7 filter-based approach to select subsets of features from among raw kinematic features. While our  
8 approach led to high performance in classifying MMH tasks and distinguishing among task  
9 conditions, the method used here – Minimal-Redundancy-Maximal-Relevancy – did not consider  
10 the temporal (time series) nature of the data. Future work should consider using feature selection  
11 methods that preserve temporal information. Third, we used machine learning algorithms that  
12 consist of non-linear structures, making them highly non-transparent in arriving at their  
13 decisions. Understanding the process or reasoning behind predictions is crucial – a concept  
14 known as *Explainable AI*. Integrating Explainable AI into machine learning can facilitate  
15 verification of predictions, systematic identification of potential flaws and biases, and the  
16 understanding of the underlying decision-making processes of machine learning algorithms  
17 (Samek et al., 2017). Fourth, we standardized some experimental aspects – such as participant  
18 clothing, lighting source and brightness, and camera positioning – which might have improved  
19 the ability to track whole-body kinematics using MMC. Thus, future work should explore the  
20 effects of varying these conditions.

21 Our tasks included several components commonly seen in the workplace that could lead to  
22 environmental occlusions, and three MMC systems were used to monitor whole-body kinematics  
23 to reduce the impact of such occlusions. Kotsifaki et al. (2018) found that increasing the number  
24 of cameras could enhance tracking of whole-body kinematics. Nevertheless, there will likely be  
25 practical constraints on the number of cameras that is feasible in practice, such as due to  
26 workflow and space restrictions. Therefore, future work should explore the possibility of using a  
27 single MMC or alternate configurations of two MMCs for physical exposure assessment.  
28 Recently, though, video surveillance cameras (plain cameras) are becoming more widely used in  
29 occupational settings (e.g., manufacturing) to enhance worker safety and to track productivity  
30 (Cocca et al., 2016; Kostal et al., 2022; Xu et al., 2015). We recommend investigating the  
31 feasibility of using plain cameras as an alternative approach to tracking body kinematics, which  
32 could be more efficient and less costly than optical or IMU systems (although the current MMC  
33 system was relatively inexpensive).

34 We removed data corresponding to idle time, since this part of the data included the calibration  
35 process and carrying the box from Task 5 to the cart, neither of which were among the MMH  
36 tasks of interest here. However, in practice, workers may perform various motions during idle  
37 times, which could be misinterpreted as MMH tasks. Future enhancements of our classification  
38 algorithms could include the ability to automatically filter out idle times and to differentiate  
39 random, non-MMH motions from actual MMH tasks. Doing so would help to ensure more robust  
40 performance, even using untrimmed, real-world video data.

## 41 4.5 Practical Applications of Our Findings

42 We suggest that the GRU model and TOP-80 feature set could be used as the base approach for  
43 practical MMH task classification using MMC. Although the GRU model generally exhibited  
44 less accuracy (by 1%) compared to the Bi-LSTM and BGRU, this marginal difference seems of

1 limited practical relevance. Further, training any of the RNN models with the TOP-80 feature set  
2 yielded a mean accuracy comparable to that using the RF features and roughly 2% better than  
3 when using the TOP-60 feature. Notably, the GRU model with the TOP-80 feature set required  
4 19 – 35% less mean epoch training time (the number of times that the model works through the  
5 entire training dataset) than when using the other models. This faster training time likely  
6 stemmed from the fact that the GRU model has a simplified architecture with two gates and  
7 fewer parameters, making it more computationally efficient for large datasets (Khandelwal et al.,  
8 2016). Moreover, using a streamlined feature set (vs. the RF feature set) helps promote a more  
9 interpretable model and contributes to improved generalization performance, by reducing the  
10 likelihood of capturing noise or irrelevant patterns in the data (Chen et al., 2017; Xue et al.,  
11 2015). Practitioners could also benefit from using a streamlined feature set, as continuous  
12 physical exposure assessments remain computationally expensive when using larger feature sets.  
13 Additionally, a streamlined feature set could inform the performance of our model when MMC is  
14 affected by occlusion and the RF feature set cannot be used.

15 Using RNN-based models and MMC is a novel approach to quantifying physical exposure  
16 assessment; as such, some comments about using this approach are warranted. One advantage is  
17 that using MMC provides a non-intrusive method for obtaining whole-body kinematics, and  
18 classification performance appears comparable to earlier results reported using inertial  
19 measurement units (IMUs). For example, earlier work that used Bi-LSTM and IMUs to classify  
20 MMH tasks reported a mean precision of 92% (Porta et al., 2021), while we achieved a mean  
21 precision of 91 – 92%. Moreover, one limitation of earlier work is that pushing and pulling tasks  
22 were often misclassified when using wearable sensors, by up to 5% (Kim & Nussbaum, 2014;  
23 Mokhlespour Esfahani, 2018; Porta et al., 2021). It is important to distinguish between pulling  
24 and pushing, though, since these tasks impose different loads on the low back (Hoozemans et al.,  
25 2004). Our approach was able to distinguish between pulling and pushing tasks with ambient  
26 sensing (i.e., MMC), with reasonable precision and recall (Figure 6). Using a sequence-to-  
27 sequence classification approach might have enhanced performance of our model, by capturing  
28 temporal dependencies and enabling RNN models to incorporate contextual information from the  
29 entire input sequence, in contrast to a sample-by-sample method (Yin et al., 2017).

## 30 **5.0 Conclusions**

31 Effectively measuring and monitoring physical exposures in the workplace is critical to assessing  
32 and controlling the risk of WMSDs. Several tools are available but are often limited in accuracy  
33 and scope, and their use can be quite resource-intensive. We evaluated the use of ambient sensors  
34 here, specifically MMC, together with machine-learning models, to classify among eight diverse  
35 MMH tasks and several specific task conditions. The classification of MMH tasks, using the  
36 MMC data, demonstrated rather robust performance across several classification models and  
37 input feature sets: mean precision, recall, and F1-score were 85 – 97%. Performance for  
38 classifying hand configuration was quite high, with mean precision, recall, and F1-scores of up  
39 to 97%. However, performance in classifying lift origin varied substantially depending on feature  
40 sets and between males and females. Overall, our findings indicate that the proposed approach  
41 has the potential to efficiently and effectively quantify MMH tasks exposures, offering a balance  
42 of simplicity and non-intrusiveness in exposure assessments. Future work will be needed,  
43 though, to assess the ability of the proposed method among diverse workers and in real working  
44 conditions.

1 **6.0 Acknowledgements**

2 Support for this study was provided by the National Safety Council (NSC). The contents of this  
3 paper are solely the responsibility of the authors and do not necessarily represent the official  
4 views of NSC. We thank Ms. Sarah Iridiastadi for her assistance in data analysis.

## 1 7.0 REFERENCES

- 2 Alawneh, L., Mohsen, B., Al-Zinati, M., Shatnawi, A., & Al-Ayyoub, M. (2020). A comparison of  
3 unidirectional and bidirectional lstm networks for human activity recognition. 2020 IEEE  
4 International Conference on Pervasive Computing and Communications Workshops (PerCom  
5 Workshops),
- 6 Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J.*  
7 *Advance Soft Compu. Appl*, 5(3), 176-204.
- 8 Andersen, J. H., Kaergaard, A., Mikkelsen, S., Jensen, U. F., Frost, P., Bonde, J. P., Fallentin, N., & Thomsen,  
9 J. F. (2003). Risk factors in the onset of neck/shoulder pain in a prospective study of workers in  
10 industrial and service companies. *Occup Environ Med*, 60(9), 649-654.  
11 <https://doi.org/10.1136/oem.60.9.649>
- 12 Antwi-Afari, M. F., Li, H., Yu, Y., & Kong, L. (2018). Wearable insole pressure system for automated  
13 detection and classification of awkward working postures in construction workers. *Automation in*  
14 *Construction*, 96, 433-441. <https://doi.org/10.1016/j.autcon.2018.10.004>
- 15 Arisoy, E., Sethy, A., Ramabhadran, B., & Chen, S. (2015). Bidirectional recurrent neural network language  
16 models for automatic speech recognition. 2015 IEEE International Conference on Acoustics,  
17 Speech and Signal Processing (ICASSP),
- 18 Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S.,  
19 Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts,  
20 taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.
- 21 Barazandeh, B., Bastani, K., Rafieisakhaei, M., Kim, S., Kong, Z., & Nussbaum, M. A. (2017). Robust sparse  
22 representation-based classification using online sensor data for monitoring manual material  
23 handling tasks. *IEEE Transactions on Automation Science and Engineering*, 15(4), 1573-1584.
- 24 Bernard, B. P., & Putz-Anderson, V. (1997). Musculoskeletal disorders and workplace factors; a critical  
25 review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper  
26 extremity, and low back.
- 27 Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn  
28 inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3), 1-33.
- 29 Burdorf, A., & van Riel, M. (1996). Design of strategies to assess lumbar posture during work.  
30 *International Journal of Industrial Ergonomics*, 18(4), 239-249.
- 31 Capela, N. A., Lemaire, E. D., & Baddour, N. (2015). Feature selection for wearable smartphone-based  
32 human activity recognition with able bodied, elderly, and stroke patients. *PLOS ONE*, 10(4),  
33 e0124414.
- 34 Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-  
35 sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
- 36 Chen, Q., Zhang, M., & Xue, B. (2017). Feature selection to improve generalization of genetic  
37 programming for high-dimensional symbolic regression. *IEEE Transactions on Evolutionary*  
38 *Computation*, 21(5), 792-806.
- 39 Ciriello, V. M., Snook, S. H., Hashemi, L., & Cotnam, J. (1999). Distributions of manual materials handling  
40 task parameters. *International Journal of Industrial Ergonomics*, 24(4), 379-388.
- 41 Cocca, P., Marciano, F., & Alberti, M. (2016). Video surveillance systems to enhance occupational safety:  
42 A case study. *Safety Science*, 84, 140-148.
- 43 da Costa, B. R., & Vieira, E. R. (2010). Risk factors for work-related musculoskeletal disorders: A  
44 systematic review of recent longitudinal studies. *Am J Ind Med*, 53(3), 285-323.  
45 <https://doi.org/10.1002/ajim.20750>

- 1 David, G. C. (2005). Ergonomic methods for assessing exposure to risk factors for work-related  
2 musculoskeletal disorders. *Occup Med (Lond)*, 55(3), 190-199.  
3 <https://doi.org/10.1093/occmed/kqi082>
- 4 Dempsey, P. G. (1999). Utilizing criteria for assessing multiple-task manual materials handling jobs.  
5 *International Journal of Industrial Ergonomics*, 24(4), 405-416. <https://doi.org/Doi>  
6 10.1016/S0169-8141(99)00007-4
- 7 Dianat, I., Molenbroek, J., & Castellucci, H. I. (2018). A review of the methodology and applications of  
8 anthropometry in ergonomics and product design. *Ergonomics*, 61(12), 1696-1720.
- 9 Djordjević, K. L., Jordović-Pavlović, M. I., Čojbašić, Ž., Galović, S., Popović, M. N., Nešić, M. V., &  
10 Markusev, D. D. (2022). Influence of data scaling and normalization on overall neural network  
11 performances in photoacoustics. *Optical and Quantum Electronics*, 54(8), 501.
- 12 Domnich, A., & Anbarjafari, G. (2021). Responsible AI: Gender bias assessment in emotion recognition.  
13 *arXiv preprint arXiv:2103.11436*.
- 14 Dwarampudi, M., & Reddy, N. (2019). Effects of padding on LSTMs and CNNs. *arXiv preprint*  
15 *arXiv:1903.07288*.
- 16 Escorcía, V., Dávila, M. A., Golparvar-Fard, M., & Niebles, J. C. (2012). Automated vision-based  
17 recognition of construction worker actions for building interior construction operations using  
18 RGBD cameras. Construction Research Congress 2012: Construction Challenges in a Flat World,
- 19 Garg, A., Chaffin, D. B., & Freivalds, A. (1982). Biomechanical stresses from manual load lifting: a static vs  
20 dynamic evaluation. *IIE Transactions*, 14(4), 272-281.
- 21 Gary, A. W., Marras, W. S., & PARNIANPouR, M. (1996). Trunk kinematics of one-handed lifting, and the  
22 effects of asymmetry and load weight. *Ergonomics*, 39(2), 322-334.
- 23 Ghezelbash, F., Eskandari, A. H., Robert-Lachaine, X., Cao, S., Pesteie, M., Qiao, Z., Shirazi-Adl, A., &  
24 Larivière, C. (2024). Machine learning applications in spine biomechanics. *Journal of*  
25 *Biomechanics*, 111967.
- 26 Golabchi, A., Han, S., & Fayek, A. R. (2016). A fuzzy logic approach to posture-based ergonomic analysis  
27 for field observation and assessment of construction manual operations. *Canadian Journal of*  
28 *Civil Engineering*, 43(4), 294-303.
- 29 Gong, J., Caldas, C. H., & Gordon, C. (2011). Learning and classifying actions of construction workers and  
30 equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Advanced*  
31 *Engineering Informatics*, 25(4), 771-782. <https://doi.org/10.1016/j.aei.2011.06.002>
- 32 Graves, A., Jaitly, N., & Mohamed, A. (2013, 8-12 Dec. 2013). Hybrid speech recognition with Deep  
33 Bidirectional LSTM. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding,
- 34 Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. 2008 Fourth  
35 international conference on natural computation,
- 36 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*, 9(8), 1735-1780.  
37 <https://doi.org/10.1162/neco.1997.9.8.1735>
- 38 Hoozemans, M. J., Kuijjer, P. P. F., Kingma, I., Van Dieën, J. H., De Vries, W. H., Van Der Woude, L. H.,  
39 Veeger, D. J., Van Der Beek, A. J., & Frings-Dresen, M. H. (2004). Mechanical loading of the low  
40 back and shoulders during pushing and pulling activities. *Ergonomics*, 47(1), 1-18.
- 41 Jordao, A., Nazare Jr, A. C., Sena, J., & Schwartz, W. R. (2018). Human activity recognition based on  
42 wearable sensor data: A standardization of the state-of-the-art. *arXiv preprint arXiv:1806.05226*.
- 43 Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network  
44 architectures. International conference on machine learning,
- 45 Khandelwal, S., Lecouteux, B., & Besacier, L. (2016). *Comparing GRU and LSTM for automatic speech*  
46 *recognition LIG*].

- 1 Khosrowpour, A., Niebles, J. C., & Golparvar-Fard, M. (2014). Vision-based workplace assessment using  
2 depth images for activity analysis of interior construction operations. *Automation in*  
3 *Construction*, 48, 74-87. <https://doi.org/10.1016/j.autcon.2014.08.003>
- 4 Kim, S., & Nussbaum, M. A. (2014). An evaluation of classification algorithms for manual material  
5 handling tasks based on data obtained using wearable technologies. *Ergonomics*, 57(7), 1040-  
6 1051. <https://doi.org/10.1080/00140139.2014.907450>
- 7 Kostal, P., Prajova, V., Vaclav, S., & Stan, S.-D. (2022). An Overview of the Practical Use of the CCTV System  
8 in a Simple Assembly in a Flexible Manufacturing System. *Applied System Innovation*, 5(3), 52.
- 9 Kotsifaki, A., Whiteley, R., & Hansen, C. (2018). Dual Kinect v2 system can capture lower limb kinematics  
10 reasonably well in a clinical setting: concurrent validity of a dual camera markerless motion  
11 capture system in professional football players. *BMJ Open Sport Exerc Med*, 4(1), e000441.  
12 <https://doi.org/10.1136/bmjsem-2018-000441>
- 13 LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (2002). Efficient backprop. In *Neural networks: Tricks of the*  
14 *trade* (pp. 9-50). Springer.
- 15 Levin, E. (1990). A recurrent neural network: Limitations and training. *Neural networks*, 3(6), 641-650.
- 16 Li, G., & Buckle, P. (1999). Current techniques for assessing physical exposure to work-related  
17 musculoskeletal risks, with emphasis on posture-based methods. *Ergonomics*, 42(5), 674-695.  
18 <https://doi.org/10.1080/001401399185388>
- 19 Liberty Mutual Insurance. (2023). *2023 Workplace Safety Index: The Top 10 Causes of Disabling Injuries*.  
20 <https://business.libertymutual.com/insights/2023-workplace-safety-index/>
- 21 Lim, S., & D'Souza, C. (2020). A narrative review on contemporary and emerging uses of inertial sensing  
22 in occupational ergonomics. *International Journal of Industrial Ergonomics*, 76, 102937.
- 23 Lindbeck, L., & Kjellberg, K. (2001). Gender differences in lifting technique. *Ergonomics*, 44(2), 202-214.
- 24 Logar, A. M., Corwin, E. M., & Oldham, W. J. (1993). A comparison of recurrent neural network learning  
25 algorithms. IEEE International Conference on Neural Networks,
- 26 Luo, H., Xiong, C., Fang, W., Love, P. E., Zhang, B., & Ouyang, X. (2018). Convolutional neural networks:  
27 Computer vision-based workforce activity assessment in construction. *Automation in*  
28 *Construction*, 94, 282-289.
- 29 Lynn, H. M., Pan, S. B., & Kim, P. (2019). A deep bidirectional GRU network model for biometric  
30 electrocardiogram classification based on recurrent neural networks. *IEEE Access*, 7, 145395-  
31 145405.
- 32 Markovitch, S., & Rosenstein, D. (2002). Feature generation using general constructor functions. *Machine*  
33 *Learning*, 49, 59-98.
- 34 Marras, W. S., Cutlip, R. G., Burt, S. E., & Waters, T. R. (2009). National occupational research agenda  
35 (NORA) future directions in occupational musculoskeletal disorder health research. *Applied*  
36 *ergonomics*, 40(1), 15-22.
- 37 Marras, W. S., & Davis, K. G. (1998). Spine loading during asymmetric lifting using one versus two hands.  
38 *Ergonomics*, 41(6), 817-834.
- 39 Martinez, R., Bouffard, J., Michaud, B., Plamondon, A., Côté, J. N., & Begon, M. (2019). Sex differences in  
40 upper limb 3D joint contributions during a lifting task. *Ergonomics*, 62(5), 682-693.
- 41 MassirisFernández, M., Fernández, J. Á., Bajo, J. M., & Delrieux, C. A. (2020). Ergonomic risk assessment  
42 based on computer vision and machine learning. *Computers & Industrial Engineering*, 149,  
43 106816.
- 44 Microsoft Corporation. (2022). *Azure Kinect body tracking joints*. [https://learn.microsoft.com/en-](https://learn.microsoft.com/en-us/azure/kinect-dk/about-azure-kinect-dk)  
45 [us/azure/kinect-dk/about-azure-kinect-dk](https://learn.microsoft.com/en-us/azure/kinect-dk/about-azure-kinect-dk)
- 46 Microsoft Corporation. (2022 ). *Azure Kinect body tracking joints*. Retrieved 10/23/2023 from  
47 <https://docs.microsoft.com/en-gb/azure/Kinect-dk/body-joints>

1 Mokhlespour Esfahani, M. I. (2018). *Development and Assessment of Smart Textile Systems for Human*  
2 *Activity Classification* [Virginia Tech].

3 Nath, N. D., Akhavian, R., & Behzadan, A. H. (2017). Ergonomic analysis of construction worker's body  
4 postures using wearable mobile sensors. *Appl Ergon*, *62*, 107-117.  
5 <https://doi.org/10.1016/j.apergo.2017.02.007>

6 Park, S., Park, J., Al-Masni, M. A., Al-Antari, M. A., Uddin, M. Z., & Kim, T.-S. (2016). A depth camera-  
7 based human activity recognition via deep learning recurrent neural network for health and  
8 social care services. *Procedia Computer Science*, *100*, 78-84.

9 Pedersen, S. J., Kitic, C. M., Bird, M.-L., Mainsbridge, C. P., & Cooley, P. D. (2016). Is self-reporting  
10 workplace activity worthwhile? Validity and reliability of occupational sitting and physical activity  
11 questionnaire in desk-based workers. *BMC Public Health*, *16*, 1-6.

12 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,  
13 Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine*  
14 *Learning research*, *12*, 2825-2830.

15 Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-  
16 dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and*  
17 *Machine Intelligence*, *27*(8), 1226-1238.

18 Plamondon, A., Larivière, C., Denis, D., Mecheri, H., Nastasia, I., & group, I. M. r. (2017). Difference  
19 between male and female workers lifting the same relative load when palletizing boxes. *Applied*  
20 *ergonomics*, *60*, 93-102.

21 Plamondon, A., Lariviere, C., Denis, D., St-Vincent, M., Delisle, A., & Group, I. M. R. (2014). Sex  
22 differences in lifting strategies during a repetitive palletizing task. *Applied Ergonomics*, *45*(6),  
23 1558-1569.

24 Plantard, P., Shum, H. P. H., Le Pierres, A. S., & Multon, F. (2017). Validation of an ergonomic assessment  
25 method using Kinect data in real workplace conditions. *Applied ergonomics*, *65*, 562-569.  
26 <https://doi.org/10.1016/j.apergo.2016.10.015>

27 Porta, M., Kim, S., Pau, M., & Nussbaum, M. A. (2021). Classifying diverse manual material handling tasks  
28 using a single wearable sensor. *Applied Ergonomics*, *93*, 103386.

29 Preece, S. J., Goulermas, J. Y., Kenney, L. P., Howard, D., Meijer, K., & Crompton, R. (2009). Activity  
30 identification using body-mounted sensors—a review of classification techniques. *Physiological*  
31 *measurement*, *30*(4), R1.

32 Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random  
33 forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

34 Punnett, L., & Wegman, D. H. (2004). Work-related musculoskeletal disorders: the epidemiologic  
35 evidence and the debate. *J Electromyogr Kinesiol*, *14*(1), 13-23.  
36 <https://doi.org/10.1016/j.jelekin.2003.09.015>

37 Rezagholi, M., Mathiassen, S. E., & Liv, P. (2012). Cost efficiency comparison of four video-based  
38 techniques for assessing upper arm postures. *Ergonomics*, *55*(3), 350-360.

39 Roberts, D., Torres Calderon, W., Tang, S., & Golparvar-Fard, M. (2020). Vision-based construction worker  
40 activity analysis informed by body posture. *Journal of Computing in Civil Engineering*, *34*(4),  
41 04020017.

42 Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding,  
43 visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

44 Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal*  
45 *Processing*, *45*(11), 2673-2681.

46 Senator, T. E. (2005). Multi-stage classification. Fifth IEEE international conference on data mining  
47 (ICDM'05),

1 Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification  
2 performance. *Applied Soft Computing*, 97, 105524.

3 Song, B., Kamal, A. T., Soto, C., Ding, C., Farrell, J. A., & Roy-Chowdhury, A. K. (2010). Tracking and activity  
4 recognition through consensus in distributed camera networks. *IEEE Trans Image Process*,  
5 19(10), 2564-2579. <https://doi.org/10.1109/TIP.2010.2052823>

6 Tang, S., & Golparvar-Fard, M. (2021). Machine Learning-Based Risk Analysis for Construction Worker  
7 Safety from Ubiquitous Site Photos and Videos. *Journal of Computing in Civil Engineering*, 35(6),  
8 04021020. <https://doi.org/Artn> 04021020

9 10.1061/(Asce)Cp.1943-5487.0000979

10 U.S. Bureau of Labor Statistics. (2021). *Nonfatal Occupational Injuries and Illnesses Requiring Days Away*  
11 *from Work*. bls.gov. Retrieved May 7 from <https://www.bls.gov/data/home.htm>

12 Wang, J., & Sainburg, R. L. (2007). The dominant and nondominant arms are specialized for stabilizing  
13 different features of task performance. *Experimental Brain Research*, 178, 565-570.

14 Wang, Y., Huang, J., He, T., & Tu, X. (2020). Dialogue intent classification with character-CNN-BGRU  
15 networks. *Multimedia Tools and Applications*, 79(7), 4553-4572.

16 Waters, T. R., Dick, R. B., Davis-Barkley, J., & Krieg, E. F. (2007). A cross-sectional study of risk factors for  
17 musculoskeletal symptoms in the workplace using data from the General Social Survey (GSS).  
18 *Journal of Occupational and Environmental Medicine*, 172-184.

19 Wells, R., Moore, A., Potvin, J., & Norman, R. (1994). Assessment of risk factors for development of work-  
20 related musculoskeletal disorders (RSI). *Appl Ergon*, 25(3), 157-164.  
21 [https://doi.org/10.1016/0003-6870\(94\)90013-2](https://doi.org/10.1016/0003-6870(94)90013-2)

22 Winter, D. A. (2009). Biomechanics and motor control of human movement. In (Third ed., pp. 49-50).  
23 John Wiley & sons.

24 Xu, X., McGorry, R. W., Chou, L.-S., Lin, J.-h., & Chang, C.-c. (2015). Accuracy of the Microsoft Kinect™ for  
25 measuring gait parameters during treadmill walking. *Gait & posture*, 42(2), 145-151.

26 Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2015). A survey on evolutionary computation approaches to  
27 feature selection. *IEEE Transactions on Evolutionary Computation*, 20(4), 606-626.

28 Yang, J., Shi, Z., & Wu, Z. (2016). Vision-based action recognition of construction workers using dense  
29 trajectories. *Advanced Engineering Informatics*, 30(3), 327-336.

30 Yang, K., Ahn, C. R., & Kim, H. (2020). Deep learning-based classification of work-related physical load  
31 levels in construction. *Advanced Engineering Informatics*, 45, 101104-101104.  
32 <https://doi.org/10.1016/j.aei.2020.101104>

33 Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language  
34 processing. *arXiv preprint arXiv:1702.01923*.

35 Yu, Y., Yang, X., Li, H., Luo, X., Guo, H., & Fang, Q. (2019). Joint-Level Vision-Based Ergonomic Assessment  
36 Tool for Construction Workers. *Journal of Construction Engineering and Management*, 145(5),  
37 04019025. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001647](https://doi.org/10.1061/(asce)co.1943-7862.0001647)

38 Zhan, K., Ramos, F., & Faux, S. (2012, 5-7 Dec. 2012). Activity recognition from a wearable camera. 2012  
39 12th International Conference on Control Automation Robotics & Vision (ICARCV),

40 Zhang, C., & Tian, Y. (2012). RGB-D camera-based daily living activity recognition. *Journal of Computer*  
41 *Vision and Image Processing*, 2(4), 12.

42 Zhang, H., Yan, X., & Li, H. (2018). Ergonomic posture recognition using 3D view-invariant features from  
43 single ordinary camera. *Automation in Construction*, 94, 1-10.  
44 <https://doi.org/10.1016/j.autcon.2018.05.033>

45 Zhou, G., Aggarwal, V., Yin, M., & Yu, D. (2022). A Computer Vision Approach for Estimating Lifting Load  
46 Contributors to Injury Risk. *IEEE Transactions on Human-Machine Systems*.

1 Zhou, X., Li, S., Liu, J., Wu, Z., & Chen, Y. F. (2024). Construction Activity Analysis of Workers Based on  
2 Human Posture Estimation Information. *Engineering*, 33, 225-236.  
3



Click here to access/download  
**Supplementary Material**  
Supplemental Material.pdf