Decomposed Physical Workload Estimation for Human-Robot Teams

Joshua Bhagat Smith, Prakash Baskaran, Julie A. Adams, Senior Member, IEEE
Collaborative Robotics and Intelligent Systems Institute
Oregon State University
Corvallis, Oregon
bhagatsj,baskarap,adamjuli@oregonstate.edu

***Abstract*—Human-robot teams operate in uncertain environments to accomplish a wide range of tasks. A dynamic understanding of the human's workload can enable fluid interactions between team members. Workload can be decomposed into workload components (e.g., cognitive, visual, speech, auditory, gross motor, fine motor, and tactile). A system that seeks to adapt interactions for a human-robot team needs to understand the distribution of workload across the different components. Prior work treated the gross motor, fine motor, and tactile components as a joint physical workload. The presented algorithm estimates gross motor, fine motor, and tactile workload for a human-robot team operating in a non-sedentary supervisory environment; however, noise and task uncertainty lead to mixed results. The metrics for this algorithm were collected using a diverse set of wearable sensors, including heart rate monitors, motion trackers, and surface electromyography sensors.**

***Index Terms*—Physical Workload, Human-Machine Teaming, Wearable Computing**

I. Introduction

Deploying human-robot teams in uncertain environments requires the robot to dynamically understand the human's internal state. This dynamic understanding must account for real-world complexities in order to enable fluid interactions between the robot and its human teammate. Workload is a measure that, in general, represents how hard a person is working, and can be decomposed into workload components (e.g., cognitive, visual, speech, auditory, gross motor, fine motor, and tactile). Incorporating accurate workload models enhances the robot's understanding of its human teammate, allowing the robot to manage the human's workload level, either through intelligent modulation of task performance, or how the robot interacts with its teammate.

Any system that seeks to adapt how it interacts with a human needs to be capable of distinguishing the specific demands placed upon the human by understanding the distribution of workload across the different components. Prior work treated the gross motor, fine motor, and tactile components as a single component, called physical workload [1]. This joint consideration only partially informs a workload estimation method, as it aggregates away key contextual information. A high physical workload may result from a high gross motor demand (e.g., lifting heavy boxes), a high fine motor demand (e.g., controlling a UAV with a joystick), or high tactile demand (e.g., commanding a fleet of robots using a tablet). Understanding the specific context of physical workload is crucial for enabling more dynamic human-robot teaming.

Future human-robot teams will be deployed in unstructured, dynamic domains (e.g., wildland fire response) unable to support environmentally embedded sensors (e.g., cameras in built structures), or complex sensor systems (e.g., EEG) that require a mostly stationary human user in a structured environment (e.g, air traffic controller). Thus, wearable physiological sensors (e.g., inertial measurement unit (IMU), electromyography (EMG)) are a core requirement for workload estimation methods, as they allow the human and the robot to act independently, enabling more flexible teaming. A modified NASA Multi-Attribute Task Battery-II (MATB-II) was used to balance data collection in a controlled environment with realistic, non-sedentary supervisory human-robot teaming.

This manuscript develops estimation models for the gross motor, fine motor, and tactile workload components. These models were trained using data from a human supervising a single remotely-located aerial robot as part of an evaluation focused on workload and task recognition.

II. Background

Humans have a limited capacity for processing information, making decisions, and dealing with physical stress. This capacity can be analyzed by viewing how humans work from a resource management perspective [2]. Workload is the ratio between the resources an individual has available to dedicate to a task and the total resources an individual has available for performing all tasks [3]. This ratio varies day-to-day and is dependent upon a range of external factors, such as expertise, emotional stress, and fatigue [4].

Workload can be decomposed into components: cognitive, visual, speech, auditory, and physical. Physical workload can be further subdivided into gross motor, fine motor, and tactile workload. *Gross motor workload* is anything that includes movement of large body segments (e.g, walking, lifting boxes), while *fine motor workload* involves more controlled movements, such as the arms and the hands. *Tactile workload* refers to tasks that require physical touch, such as typing, clicking buttons on a computer mouse, or feeling surfaces.

High and low workload values have different implications on a human's internal state. A high workload value, overload, occurs when a task requires a large amount of resources, but

978-1-6654-5238-0/22/$31.00 ©2022 IEEE

an individual only has a small amount of resources available. A low workload value, underload, occurs when a task requires relatively few resources and the individual has a large amount of available resources. The thresholds for overload and underload vary between people and tasks. Normal load is defined as the region between these two thresholds. The underload condition has been difficult to detect but presents a problem of equal magnitude to overload. People tend to become unengaged in their work during periods of underload, which leads to reduced alertness and lowered situational awareness [4].

Prior work either focused on estimating overall workload for physical tasks [5], or physical workload as a single component [1] [6]. Manjarres et al. segmented five heart-related metrics (i.e., absolute cardiac cost, relative cardiac cost, max heart rate, mean heart rate, heart-rate acceleration) into intervals that are summed to create a Frimat's score [6], which was used to construct intervals representing physical workload and task difficulty. A Random Forest model classified these heart-related metrics into physical workload intervals. Heard et al. developed neural network-based models that used wearable sensors to estimate cognitive, auditory, speech, physical, and overall workload [7]. Workload values for the experimental tasks were derived from the human performance modeling tool IMPRINT PRO [8]. The physical workload relied solely on physiological metrics (e.g., postural load, respiration rate, heart rate). These views of physical workload do not investigate the specific contributions of the gross motor, fine motor, and tactile components to overall workload.

PHYSIOPRINT is a workload assessment tool that estimates all seven workload components, including gross motor, fine motor, and tactile [9]. This tool uses electroencephalogram (EEG) and electrocardiogram (ECG) metrics to perform atomic task recognition, where the atomic tasks correspond to known tasks within IMPRINT PRO with an assigned static workload value. This approach takes a more granular approach to physical workload estimation, but the reliance on these static values overlooks the impact of individual differences.

There has been increasing interest in the use of wearable sensors for performing physical task recognition [10]. Both traditional machine learning [11] and deep learning [12] have proven effective at performing gross motor task recognition. Similarly, deep learning methods exist for recognizing fine motor tasks [13] and tactile tasks [14]. The problems of physical task recognition and physical workload estimation are distinct but share core commonalities. The success of these methods suggests that using wearable sensors for gross motor, fine motor, and tactile workload estimation holds promise.

III. Methods

The research questions focus on if the workload models can (1) reliably identify an individual's workload level (i.e., UL, NL, OL), and (2) produce a precise workload value. Identifying the relative workload level can help an adaptive system understand the nature of a human's needs, whether they need to be more engaged or require assistance. Precise workload values add additional context, by quantifying the proximity to the OL or UL threshold. An evaluation designed to estimate workload and detect tasks (part of another effort) used the NASA MATB-II [15]. The evaluation manipulated tasks, task density (i.e., workload), and task density ordering as independent variables. The task density variable (i.e., workload levels) manipulated the number of tasks initiated during a specific time period. The workload was elicited by increasing and decreasing the NASA MATB-II tasks' frequency Underload (UL), Normal load (NL), and Overload (OL). The task density ordering (i.e., workload ordering) ensured that each task density transition (e.g., UL-NL, OL-UL) occurred exactly once. Participants completed a single 52.5-minute trial using an adapted NASA MATB-II version, where the trial consisted of seven consecutive 7.5-minute task density conditions. Three task density orderings:

- O1: UL-NL-OL-UL-OL-NL-UL
- O2: NL-OL-UL-OL-NL-UL-NL
- O3: OL-UL-OL-NL-UL-NL-OL

In-situ workload ratings of each workload component were collected every 7.5 minutes, subjectively on a scale of 1 to 5.

A. Supervisory Task Environment

The supervisory task environment consisted of a modified version of the NASA MATB-II [15] and required a human to supervise a simulated aerial robot. The NASA MATB-II consists of four tasks: tracking, system management, resource management, and communication. The communication task was split into two separate tasks, communication, and communication response, in order to model the communication itself and any verbal response.

Participants were not required to move when using the original NASA MATB-II system, but real-life human-robot teaming scenarios can require movement throughout the environment. The NASA MATB-II was modified to physically separate each NASA MATB-II task; thus, requiring participants to walk between two sets of tasks, called the walking task. This physical layout is depicted in Figure 1. The modified NASA MATB-II provides control over the task environment. Each NASA MATB-II task had a separate dedicated monitor, each stationed such that the participant was unable to visually see more than two tasks simultaneously. This visual hindrance ensured that participants walked around the environment to complete the overall mission. The required equipment (e.g.,

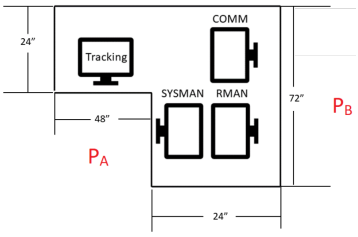
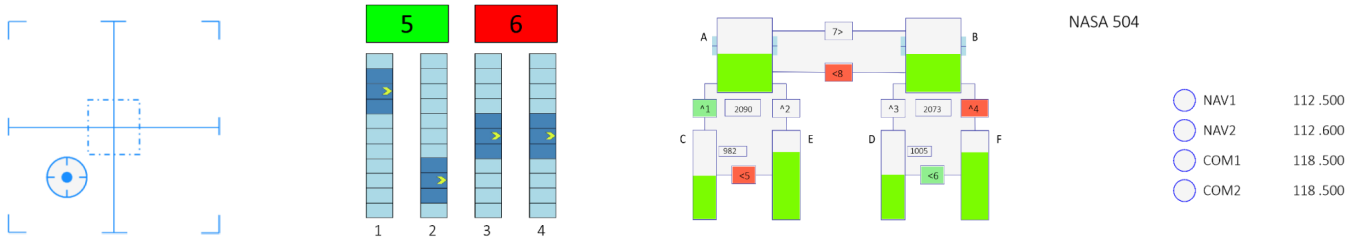


Fig. 1. The Modified NASA MATB-II physical layout. P_A and P_B represent the area participants walked between. SYSMAN: System Monitor station. RMAN: Resource Management station. COMM: Communications station.



(a) Tracking (b) System Monitoring (c) Resource Management (d) Communications

Fig. 2. The NASA MATB-II Tasks.

joystick or keyboard) to complete each task was placed in front of the respective computer monitor.

The tracking task, depicted in Figure 2a, required participants to keep the circle with a blue dot in the middle of the cross-hairs using a joystick. The underload condition consisted of a single 45-second session of manual tracking, otherwise it was automated. The overload condition required two 12-second manual tracking sessions every minute, while the normal load required one 20-second session every minute.

The system monitoring task, shown in Figure 2b, required monitoring two colored lights and four gauges. If the green (L5) or the red light (L6) turned on, the value was out of range and required resetting. The four gauges had a randomly moving up and down indicator that typically remained in the middle. Participants reset a gauge if it was out of range (i.e., too high or too low). These items were reset by pressing the corresponding number key on the keyboard's top row. The underload condition had only one out-of-range instance in the entire 7.5-minute session, overload had fifteen instances per minute, and normal load had five instances per minute.

The resource management task included six fuel tanks (A-F) and eight fuel pumps (1-8), shown in Figure 2c. The arrow by the fuel pump's number indicated the fuel flow direction. Participants were to maintain the fuel levels of Tanks A and B by turning the fuel pumps on or off. Fuel Tanks C and D had finite fuel levels, while Tanks E and F had an infinite fuel supply. A pump turned red when it was unable to pump fuel. The underload condition had 2 minutes of manual management with zero pumps failing, otherwise it was automated. The entire overload condition required manual management, with two or more pumps failing, while only the last 3.5 minutes of the normal load condition required manual management, with at most two pumps failing every minute.

The communications task required listening to air-traffic control requests for radio changes. The communication request was similar to this: "NASA 504, please change your COM 1 radio to frequency 127.550." The original MATB communications task required no speech, but a required verbal response was added. An example response is: "This is NASA 504 tuning my COM 1 radio to frequency 127.550." Participants changed the specified radio to the specified frequency by selecting the desired radio and using arrows to change the radio's frequency, as depicted in Figure 2d. Communications not directed to the participant's aircraft, indicated by the call sign, were to be ignored. The underload condition contained a single communication request with one communication response task, overload contained three communication requests with at least two communication response tasks every minute, while normal load contained up to two auditory communication requests, with only one communication response task per minute.

Finally, participants were required to walk around the tables to the other stations (e.g., from P_A to P_B in Figure 1) whenever a ping sound occurred. Participants were free to move between the tasks at any time, but the ping sound enforced a mandatory transition to the other workstations. The underload condition contained two walk requests, the overload condition incorporated seven requests per minute, and normal load had two requests per minute. Task timings and occurrences were chosen, such that the correct workload condition, or task density, was elicited. The IMPRINT Pro tool was used to model the tasks for each workload level and ordering prior to conducting the evaluation. The IMPRINT Pro tool provided anchors to choose the workload difficulty value [8]. The anchor values are not normalized across components; thus, the association between a task and the workload value varies significantly across the components.

A tutorial video described the NASA MATB-II tasks and how to accomplish the tasks. The tutorial video was followed by a 10-minute training session, during which participants gained familiarity with the task environment The training session cycled through the tasks, with each task occurring continuously for a 1-minute period, repeating the cycle one additional time. The 52.5-minute trial commenced after the training. The tasks switched rapidly and sometimes overlapped to emulate real-world scenarios. Forty-three participants (26 male, 17 female, and 2 non-binary) completed the experiment, with ages ranging from 18 to 60.

B. Feature Extraction

Objective physiological metrics were collected using the BioPac Bioharness™ BT [17], the Xsens MTw Awinda [16], and two Myo armbands [18]. The Bioharness™ collects various physiological metrics, including heart rate, respiration rate, and postural magnitude, at a 1 Hz sampling frequency. The Xsens MTw Awinda measures the participant's body pose using seventeen IMUs located on the body (see Figure 3) at 40 Hz. The two Myo armbands each measure an 8-channel surface EMG and a forearm inertial metrics at 100 Hz. Each

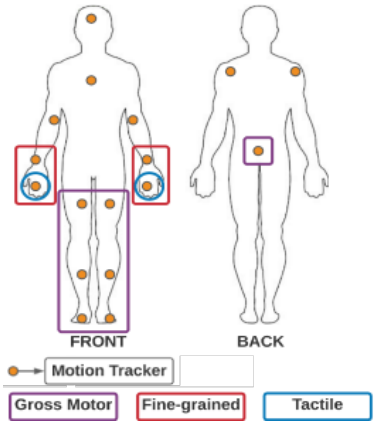


Fig. 3. Xsens IMU position on the body, with highlights for the sensors, used to estimate gross motor, fine motor, and tactile workload [16].

physical workload subcomponent relies on a different set of sensors, as different metrics correlate with each component.

The gross motor workload model combines the Bioharness' heart rate, respiration rate, and postural magnitude with the Xsens' lower body inertial metrics. Four time-based features are derived from the Bioharness^{TM} metric: mean, standard deviation (stdev.), slope, and gradient. Prior work established that these features are effective when estimating physical workload [7] and that they correlate with changes in gross motor workload [4]. The primary gross motor task in this experiment was walking. Thus, seven IMUs on the lower body (i.e., pelvis, thighs, calves, and ankles, as shown in Figure 3) produce positional and orientation measurements in three dimensions, which were used to calculate the magnitude of the acceleration and angular velocity vectors. Additionally, seven time-based features were derived using these vectors: mean, stdev., skew, kurtosis, minimum, maximum, and median.

The fine motor workload model combines the inertial data from the Xsens motion trackers on the wrists and hands, see Figure 3, with the forearm inertial and 8-channel sEMG data obtained from the two Myos. The tactile workload model also combines the inertial data from the Xsens with the EMG data from the Myo, but only incorporates the left and right-hand inertial metrics and the 8-channel surface EMG metrics for both forearms. Both models' metrics were processed in a similar manner to the gross motor model metrics. The inertial data was used to calculate the magnitude of the acceleration and angular velocity vectors in three dimensions, and the seven time-based features were extracted for inertial and EMG data.

Prior work established that the Xsens' inertial data is not prone to either significant time delay (i.e., less than 1 second), or significant amounts of drift [16] [19], which was primarily attributed to its magnetometer-based correction and a Kalman Filter. The Myo armbands use similar techniques, but no studies have verified they are not prone to time delays or drift.

All metrics were processed using a sliding window of 10 seconds. Data within the window was standardized by subtracting the mean (μ) and dividing by the standard deviation (σ), as described in Equation 1.

$$y = \frac{x - \mu}{\sigma} \tag{1}$$

C. Model Development and Validation

Neural networks have been effective at estimating cognitive, auditory, and physical workload [7]. However, recently tree-based models generally outperform on tabular datasets [20]. Thus, both a fully-connected neural network and a gradient boosting regression tree (GB Tree) were evaluated for estimating the physical workload subcomponents. The neural network was implemented using Pytorch and had 5-layer hidden layers with 256 nodes each. The GB Tree was implemented using Scikit-learn's Histogram-Based Gradient Boosting algorithm [21]. Early stopping was employed in both models to prevent overfitting. All machine learning models were validated using leave-one-subject-out (LOSO) cross validation, where the root mean squared error (RMSE) is reported by training the model on all, but one participant's data and validating using the left-out participant's data.

IV. Results

The RMSEs for the LOSO cross validation by model type are provided in Table 1. Prior work suggested that a RMSE $\leq$ 5% of the maximum workload value is sufficient to identify workload levels and produce precise workload values [4]. The gross motor target RMSE (i.e, 5% of the maximum) is 0.15, as the maximum gross motor workload value is 3. The gross motor models came the closest to the $\leq$ 5% goal, with a RMSE for both models of 0.19 (i.e., 6% of the maximum). The target for fine motor workload is 0.92, as it has a maximum value of 18.4. The fine motor GB Tree and the Neural Net resulted in a similar RMSE, approximately 10% of the fine motor workload maximum. Tactile has a maximum of 12.0; therefore, the target RMSE is 0.60. The GB Tree tactile model achieved a RMSE of 1.96, while the Neural Net had a RMSE equal to 2.08, both of which are approximately 15% of the maximum tactile workload value. Given the GB Tree's lower RMSE in all instances, it was used for the remaining analysis.

The GB Tree models were used to estimate each participant's workload for the entire experiment. Mean estimated workload was calculated every 7.5 minutes (i.e., workload interval) at 1 Hz, for each participant, resulting in an overall mean value for UL, NL, and OL workload. The medians of the resulting condition means, for all participants, and each workload component are shown in Figure 4. The figure shows that the models are unable to distinguish the relative workload

TABLE I
Workload RMSE

Component	Neural Net	GB Tree			
		Overall	O1	O2	O3
Gross Motor	0.199	0.194	0.185	0.191	0.207
Fine Motor	2.062	1.942	1.857	2.059	1.916
Tactile	2.084	1.962	1.885	2.060	1.936

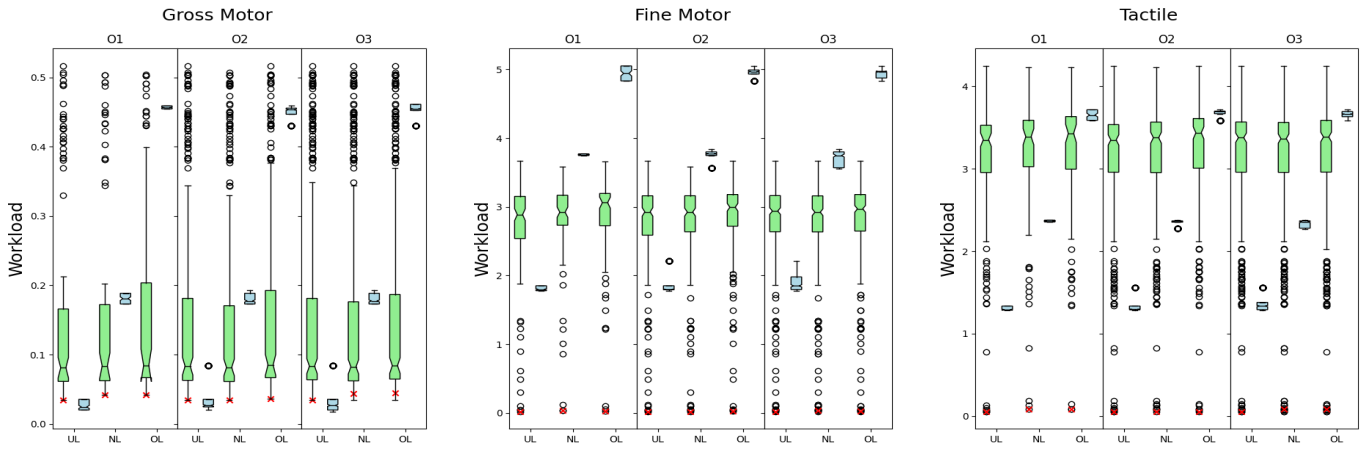


Fig. 4. Quantile plots of (a) gross motor, (b) fine motor, and (c) tactile workload by ordering for mean (green) and IMPRINT PRO (blue) estimated workloads, with the additional of outliers (white circles) and the modes (red x's).

TABLE II
IN-SITUATION SUBJECTIVE WORKLOAD VALUES ACROSS WORKLOAD ORDERINGS AND CONDITIONS.

		Workload		
Ordering	**Component**	**UL**	**NL**	**OL**
O1	Gross Motor	1.76 (0.87)	3.03 (1.09)	3.6 (1.13)
	Fine Motor	2.07 (0.9)	2.76 (0.91)	3.3 (0.84)
	Tactile	2.26 (1.08)	2.79 (1.08)	3.3 (1.26)
O2	Gross Motor	2.3 (1.05)	2.0 (1.04)	1.87 (0.81)
	Fine Motor	2.42 (1.06)	2.36 (1.25)	2.22 (1.17)
	Tactile	2.67 (1.11)	2.40 (1.08)	2.22 (1.17)
O3	Gross Motor	2.9 (1.07)	2.04 (0.66)	2.35 (1.16)
	Fine Motor	3.1 (1.02)	2.08 (0.84)	2.42 (1.27)
	Tactile	2.97 (1.11)	2.35 (1.09)	2.46 (1.3)

level for any of the three components, even for gross motor that had a RMSE close to the 5% target.

The O1 estimates follow the appropriate trend for all components, where UL $<$ NL $<$ OL. However, the difference between median values across workload levels is small. The other task density orders do not exhibit this trend. Additionally, the RMSE values were lower for the O1 participants than O2 and O3. Interestingly, this pattern matches the in-situ ratings, provided in Table II. O1 clearly shows increasing ratings across the conditions, while the OL condition had the lowest value for O2. Oddly, UL had the highest ratings for both O2 and O3. However, the mean subjective workload for UL, NL, and OL, for both O2 and O3, are all within one standard deviation of each other. Considerable noise exists in the workload estimates. There are numerous outliers across the workload components, orderings, and conditions as shown in Figure 4. Further, the workload interval estimates' mode is equal to the minimum estimate in most instances.

V. DISCUSSION

The gross motor model achieved a RMSE of 6% of the maximum workload, but neither the tactile nor fine motor models came close to the $\leq$5% goal. None of the workload models were able to reliably identify the workload level. The low RMSE values suggest that the machine learning models found a small relationship between the participants' workload and physiological metrics, but it was not strong enough to identify workload levels in arbitrary situations. The poor performance is caused by (1) noise and (2) task uncertainty.

The noise is clearly indicated by the number of outliers in the workload estimates. The mode equalling the minimum suggests noise comes from periods with low amounts of physical work, regardless of the physical workload subcomponent. Different tasks require varying amounts of physical movement. An over-representation of tasks that do not contain a lot of physical movement can introduce bias into the training process. Balancing the training dataset to account for task types and component-specific workload levels will help to improve performance. Further, the experimental design only modulates overall workload. There may be times when physical workload is low, even though overall workload is high.

High task uncertainty will exist but complicates machine learning model development. Verifying participants execute tasks in the modeled order, at the expected time, is non-trivial and highly uncertain. Task uncertainty can cause misalignment between the objective workload estimation, and the IMPRINT Pro workload model, leading to poor performance during the training process. Further, superfluous activity by the participant may have introduced additional errors. Participants were not prevented from checking the stations during the UL and NL conditions. Further, O2 and O3 introduce the overload condition earlier than O1. Participants transitioning from OL to UL, or NL may anticipate the high task volume will continue, causing them to engage in additional physical movement. This unnecessary work exacerbates the misalignment between the estimates and the IMPRINT PRO model.

One method of reducing task uncertainty is to manually label data via video coding, and only include data for active task engagement. Coding is time-consuming, not generalizable, and

will not work in real-time. Some level of uncertainty is necessary, otherwise the machine learning model will be unable to handle real-world situations. Another approach incorporates a metric for task composition. Understanding each workload component's weighted contribution to a task facilitates the machine learning model weighting the physiological signals accordingly. Thus, task composition mitigates the influence of noise from the physiological metrics. Task composition can be derived from an external model (e.g., IMPRINT PRO) or a task recognition algorithm, but external models do not account for new tasks, and task recognition is an open problem.

Both the in-situ workload ratings and the objective workload models poorly reflected the relative workload levels in O2 and O3. The two differences between the orderings are (1) the initial workload level, and (2) the workload level transitions. It is unclear whether these factors altered the participants' perceived workload, or impacted their ability to accurately report workload. O1 monotonically increases workload over the first three intervals, while O2 and O3 participants encountered OL prior to UL. These factors may have altered participants' expectations of future workload and the intended meaning of the participants' in-situ workload ratings. However, it is also possible that the in-situ workload ratings' variance is too great to make definitive conclusions. These results warrant further investigation to understand how the timing of the workload levels impacts participants' perceived workload.

Participant handedness, which was not recorded, may also impact the results. It is likely the majority of participants are right-handed, and the resulting models may be biased as a result. Further, the left-handed participants' metrics will add noise during training. This uninformed handedness mixing likely introduced additional errors.

Many human-robot teaming domains required a period of training, to allow the human to become familiar with the robot. Future work can utilize such a training period to gather demographic information (e.g., handedness), contextual information, expected tasks composition, and other individual differences that will provide more reliable training data.

VI. Conclusion

Models to estimate gross motor, fine motor, and tactile workload were developed using physiological metrics extracted from wearable sensors. These models achieved a low RMSE, but are not reliable enough to detect the desired workload component's levels. While these metrics hold promise, further investigation is needed. High task uncertainty and superfluous participant movement appear to have added significant noise to the metrics, which impacted the machine learning process. Future work will investigate methods for mitigating noise in the physiological data and incorporating contextual information, such as handedness and task composition, into machine learning models to improve performance.

Acknowledgments

The graduate students have been supported by an Office of Naval Research award N00014-21-1-2190. The contents are those of the authors and do not represent the official views of, nor an endorsement, by the Office of Naval Research.

References

[1] C. E. Harriott, T. Zhang, and J. A. Adams, "Assessing physical workload for human–robot peer-based teams," *International Journal of Human-Computer Studies*, vol. 71, no. 7-8, pp. 821–837, 2013.

[2] H. Dreyfuss, *Designing For People*. Skyhorse Publishing Inc., 2003.

[3] C. D. Wickens, S. E. Gordon, Y. Liu, and J. Lee, *An Introduction to Human Factors Engineering*. Pearson Prentice Hall Upper Saddle River, NJ, 2004, vol. 2.

[4] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, 2018.

[5] A. H. Memar and E. T. Esfahani, "Objective assessment of human workload in physical human-robot cooperation using brain monitoring," *ACM Transactions on Human-Robot Interaction*, vol. 9, no. 2, pp. 1–21, 2019.

[6] J. Manjarres, P. Narvaez, K. Gasser, W. Percybrooks, and M. Pardo, "Physical workload tracking using human activity recognition with wearable devices," *Sensors*, vol. 20, no. 1, p. 39, 2019.

[7] J. Heard and J. A. Adams, "Multi-dimensional human workload assessment for supervisory human–machine teams," *Journal of Cognitive Engineering and Decision Making*, vol. 13, no. 3, pp. 146–170, 2019.

[8] D. K. Mitchell, "Mental workload and ARL workload modeling tools," Army Research Lab Aberdeen Proving Ground MD, Tech. Rep., 2000.

[9] D. Popovic, M. Stikic, T. Rosenthal, D. Klyde, and T. Schnell, "Sensitive, diagnostic and multifaceted mental workload classifier (physioprint)," in *International Conference on Augmented Cognition*, 2015, pp. 101–111.

[10] A. Kristoffersson and M. Lindén, "A systematic review of wearable sensors for monitoring physical activity," *Sensors*, vol. 22, no. 2, p. 573, 2022.

[11] H. Allahbakhshi, L. Conrow, B. Naimi, and R. Weibel, "Using accelerometer and gps data for real-life physical activity type detection," *Sensors*, vol. 20, no. 3, p. 588, 2020.

[12] H. Gjoreski, J. Bizjak, M. Gjoreski, and M. Gams, "Comparing deep and classical machine learning methods for human activity recognition using wrist accelerometer," in *Workshop on Deep Learning for Artificial Intelligence at the International Joint Conference on Artificial Intelligence*, vol. 10, 2016, p. 970.

[13] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, "Deep residual bidir-lstm for human activity recognition using wearable sensors," *Mathematical Problems in Engineering*, 2018.

[14] E. Rahimian, S. Zabihi, A. Asif, and A. Mohammadi, "Hybrid deep neural networks for sparse surface emg-based hand gesture recognition," in *Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2020, pp. 371–374.

[15] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide," Tech. Rep., 2011.

[16] M. Paulich, M. Schepers, N. Rudigkeit, and G. Bellusci, "Xsens mtw awinda: Miniature wireless inertial-magnetic motion tracker for highly accurate 3d kinematic applications," *Xsens: Enschede, The Netherlands*, pp. 1–9, 2018.

[17] J. A. Johnstone, P. A. Ford, G. Hughes, T. Watson, and A. T. Garrett, "Bioharness™ multivariable monitoring device: part. i: validity," *Journal of Sports Science & Medicine*, vol. 11, no. 3, p. 400, 2012.

[18] M. Sathiyanarayanan and S. Rajan, "Myo armband for physiotherapy healthcare: A case study using gesture recognition application," in *IEEE International Conference on Communication Systems and Networks*, 2016, pp. 1–6.

[19] M. Al Borno, J. O'Day, V. Ibarra, J. Dunne, A. Seth, A. Habib, C. Ong, J. Hicks, S. Uhlrich, and S. Delp, "Opensense: An open-source toolbox for inertial-measurement-unit-based measurement of lower extremity kinematics over long durations," *Journal of Neuroengineering and Rehabilitation*, vol. 19, no. 1, pp. 1–11, 2022.

[20] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" *arXiv preprint arXiv:2207.08815*, 2022.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.