Predictive Model for Workload in Remote Operators During sUAS Contingency Scenarios*

Meghan Chandarana[†‡], Megan Shyr[†§], and Terence L. Tyson[¶]
NASA Ames Research Center, Moffett Field, CA 94035, USA

The increase in automated capabilities of small Uncrewed Aerial Systems (sUAS) has enabled the human operators to manage larger numbers of vehicles simultaneously. As this happens, the operational paradigm shifts to an m:N configuration where multiple operators (m) are managing multiple vehicles (N) together. However, many questions about how operators will interact with each other and share interaction across the vehicle pool are yet unanswered. Therefore, stakeholders from government and industry have partnered to develop ground control station concepts for such operations. The work presented in this paper aims to identify factors that contribute to operator workload. A supervised machine learning-based method built using Support Vector Machines and K-fold cross-validation was used to create workload prediction models for various NASA TLX subscales by leveraging features related to interactions and their relative timings during *m:N* operations. Results show that the models yielded fairly high predictive accuracies ranging from ∼60-75%.

I. Introduction

Currently, stakeholders from government and private industry are interested in using small Uncrewed Aerial Systems (sUAS) for applications across a variety of domains such as search and rescue [1, 2], power line inspection [3, 4], package delivery [5, 6] and military operations [7]. With the increase in automated capabilities of sUAS, human operators are now able to simultaneously manage larger numbers of vehicles. Often, within one region, multiple operators (each with their own set of vehicles to manage) are cooperating together. This paradigm where multiple operators are managing multiple vehicles is referred to as *m:N* operations, where m is the number of operators and N is the number of vehicles being managed [8].

Within the *m:N* operational paradigm, many questions remain unanswered about how remote operators can maintain situational awareness of highly automated vehicles, vehicles' state, and vehicles' surroundings such that they are able to effectively intervene when necessary to reroute vehicles in off-nominal situations. As part of NASA's Transformational Tools and Technologies (TTT) project, work is aimed at developing and testing various ground control station (GCS) concepts for pilots who would be responsible for a fleet of sUAS in a UAS Traffice Management (UTM) context [9].

As part of the NASA-Joby partnership, our previous work involved conducting a user study to examine the feasibility and effectiveness of a GCS interface concept that enables remote operators to direct multiple sUAS reroutes during contingency situations [8, 10]. The work presented in this paper utilizes the data collected during that study to develop a supervised machine learning-based prediction method for subjective workload in remote operators given various performance parameters associated with sUAS reroutes.

II. Related Work

A. Supervisory Control

Supervisory control refers to when a human utilizes a computer and various displays to intermittently communicate and send commands to a controlled process or task environment that implements the commands through sensors

*This work was supported by NASA's Autonomous Systems Enduring Discipline under the Transformational Tools and Technologies Revolutionary Aviation Mobility Subproject. The views expressed are those of the authors and do not necessarily reflect the official policy or position of NASA or the U.S. Government.

†These authors contributed equally.

‡Computer Engineer, Intelligent Systems Division, AIAA Member
§Research Student Trainee, Human Systems Integration Division
¶Research AST Human/Machine Systems, Human Systems Integration Division

and actuators [11, 12]. In the context of remote operations of automated agents, human supervisory control involves monitoring and intervening tasks related to the automated vehicles under their command. Feedback from sensors communicating with the interface used by the human enables them to maintain the appropriate level of situational awareness [13–15]. The supervisory control paradigm for control of multiple, automated agents spans across application domains including space [16, 17], search and rescue [2, 18], warehouse automation [19], power management [20], swarms [1, 21], and many other areas.

B. Workload Prediction

Previous workload prediction methods have explored real-time workload prediction during supervisory control tasks by tracking operators' attention [22] and amount of time the operator was occupied [23]. In addition, operator capacity to manage multiple vehicles has been shown to be related to the workload they felt through the amount of time it took for them to cognitively reorient themselves [14]. Cummings et al.'s work related operator performance to mission costs and complexity, and showed that both were able to predict operator performance [24]. Although these previous methods explored the effects of mission parameters and operator attention on operator performance and workload, none have explored the impact of the number of operator interactions and their initiating time relative to contingency onsets on workload. Other previous work has relied on extensive task analysis to identify all tasks expected to be performed by the operator [25].

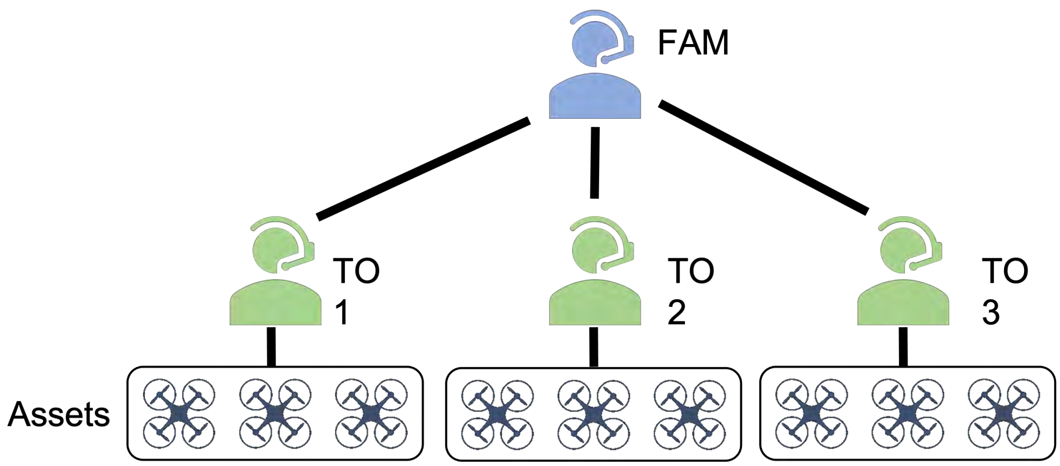

Fig. 1 Diagram of relationship between assets, Tactical Operators (TOs) and the Fleet Area Manager (FAM).

III. Study Overview

Twelve pilots participated in the study. All were Part 107 certified and/or a US Air Force pilot. For the duration of the study, subjects took the role of a remote GCS operator – Tactical Operator (TO) – tasked with supervising highly automated sUAS food delivery operations taking place in the San Diego, CA area while being subject to UAS Volume Reservation (UVR) events. These volumes of reserved airspace presented as temporary no-fly zones that spanned all altitudes, resulting from environment factors (e.g., airspace restriction from the local fire department). The study assumed that multiple TOs were managing their own sets of assigned assets under the supervision of a Fleet Area Manager (Fig. 1). Leveraging the UTM operational framework, the airspace consisted of 8 corridors, 2 hub locations, and 4 delivery sites (Fig. 2). The study employed a 2 X 2 within-subjects design where each subject completed 4 trials simulating a flight operation with a 1:12 subject-to-vehicle ratio. The study design manipulated the number of vehicles affected by each UVR (2 vs. 4 vehicles) and automation support (manual vs. assisted) in a given trial. Due to inconsistencies found in the scenario design, modifications were necessary after the first 2 subjects. Therefore, data from only the last 10 subjects were used in data analysis and the development of the prediction model presented below.

The trial order was blocked by automation condition and counterbalanced across subjects. Each trial lasted about 15 minutes and simulated 2 separate UVRs across different corridors. These caused portions of the flight corridor to become temporarily unavailable. As a result, subjects were required to reroute all affected vehicles around each UVR.

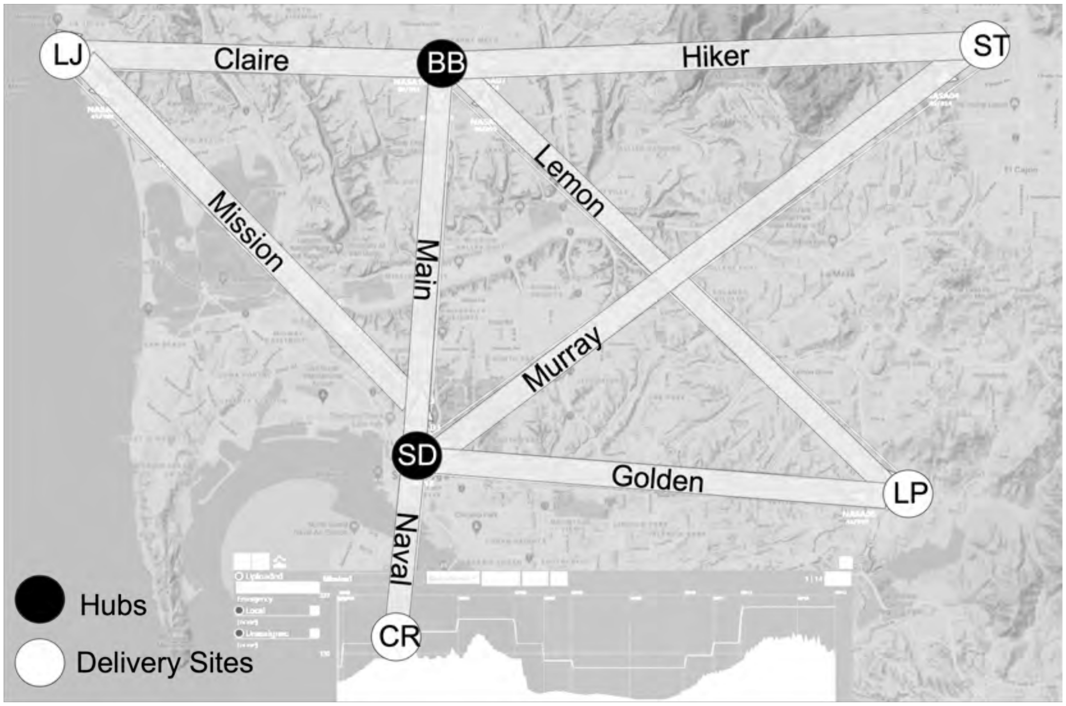

Fig. 2 Airspace and corridors used for the study. Black circles indicate hubs and white circles indicate delivery sites. Each has unique initials. Corridors, each with a unique name, are depicted with grey rectangles [8].

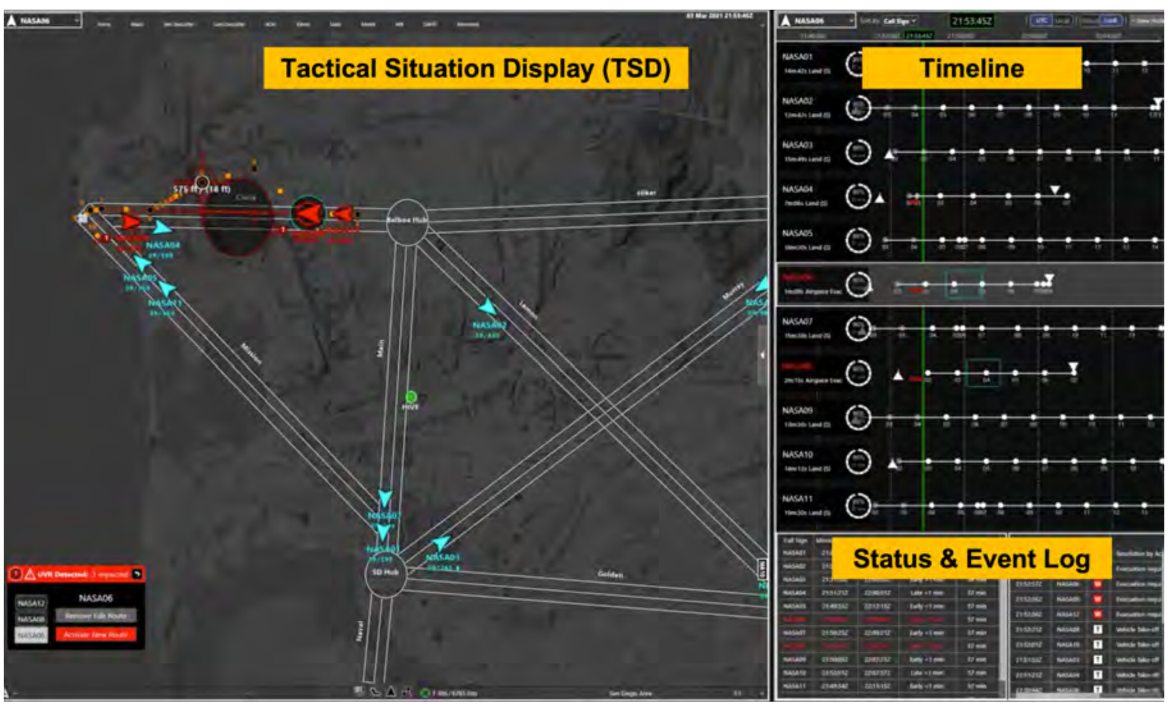

Fig. 3 Screen capture of the GCS during an experimental trial. The Tactical Situation Display (TSD) on the left depicts vehicle information, navigation controls, and airspace information. The Timeline display on the top right shows future major events for each vehicle while the Status & Event Log on the bottom right gives previous events and status messages.

Fig. 3 shows an example of what the GCS would show the operator if a UVR was present. The UVR is shown as a dark gray circle in the Claire corridor on the Tactical Situation Display (TSD) on the left. In the given example, the operator has selected a waypoint to modify. The selected waypoint is surrounded by a red crosshair and amber, dashed route segments, indicating that the proposed waypoint has a terrain conflict at the current altitude. The pop-up message in the bottom left of the TSD is known as a Toast Message and appears in the assisted automation support condition trials. The Toast Message provides an automated reroute for all vehicles affected by the UVR. The timeline display on the right shows the planned waypoints for each vehicle. The Status & Event Log at the bottom right of the interface shows important events for each vehicle. Vehicles with flight paths that currently still conflict with the UVR are shown in red on the TSD, Timeline and Status & Event Log.

In addition to the within-subjects design described above, an additional manipulation introduced a terrain conflict that required altitude adjustments when creating reroutes for 25% of the vehicles affected by the UVRs. More specifically, 1 out of the 2 UVRs was close to terrain conflicts where only half of the vehicles affected by the UVR would be affected by the terrain. In the assisted automation condition suggested reroutes were provided for each of the vehicles affected by the UVR; however, these reroute suggestions did not account for necessary altitude corrections due to terrain conflicts.

A. Subject-Surrogate Data Collection Paradigm

Due to the public health restrictions placed on on-site work during the COVID-19 pandemic, this study utilized a subject-surrogate paradigm for data collection. This paradigm required that subjects (i.e., remote pilots) relay all commands verbally to a researcher who then carried out the actions on the GCS. Subjects were not able to have direct control of the GCS themselves, but they were able to view a real-time video feed (through Microsoft Teams) of the GCS during the entire length of the trial to watch as events unfolded and as the surrogate carried out the requested actions. This data collection method required the development of a verbal protocol that paired command actions with display objects to standardize and simplify interactions between the subject and surrogate to help reduce delays and frustration. Although the verbal protocol was not ideal, it enabled real-time data acquisition with a subject in the loop while also adhering to the COVID-19 restrictions.

B. Data Collected

During each trial, various performance metrics related to the defined reroutes and the efficiency in defining them were collected for each vehicle affected by the UVR. The metrics obtained are given in Section IV.B. The NASA Task Load Index (TLX) was administered at the conclusion of each trial to obtain subjective workload ratings across 6 subscales: mental demand, temporal demand, physical demand, effort, performance, and frustration level [26]. One critical limitation of the surrogate-subject team was in the relaying of verbal commands communicated from the subject to the surrogate, effectively distributing the workload across the two agents. Due to the nature of this paradigm, we collected NASA TLX scores from both the subject and the surrogate researcher individually after each experimental trial. For the prediction model development outlined below, automation support level was not used to distinguish data, and data from UVRs with terrain conflicts were not included in the data set.

IV. Workload Prediction Model

A linear kernel was used to train a Support Vector Machine (SVM) classifier to predict each workload measure from the NASA TLX subscale for subjects. The features used for training the models were selected from seven performance metrics (refer to Section IV.B) collected during the study outlined above. A K-Fold Cross-Validation method was used to systematically identify which feature combination and respective number of K-Folds (f) produced the most accurate classifier for each NASA TLX subscale. All possible combinations of features were evaluated for a range of K-Folds ($f \in \{2, 3, ..., 10\}$). Surrogate data was not used for building the final models described below.

The analyses were run using the *scikit-learn* Python module [27]. To avoid overfitting our data, the data was randomly distributed into "folds" for the training and testing datasets 100 different times for each <feature combination, K-Fold> pair. Average results across the 100 different randomly chosen "folds" were used to compare against other <feature combination, K-Fold> pairs. Fig. 4 outlines our approach.

A. Binary Classifier

The NASA TLX includes 6 individual subscales. Each subscale was binned using a median split (if score < median, assign 1; else 0). The number of vehicles affected was binned as shown in the table where cases with 4 vehicles were

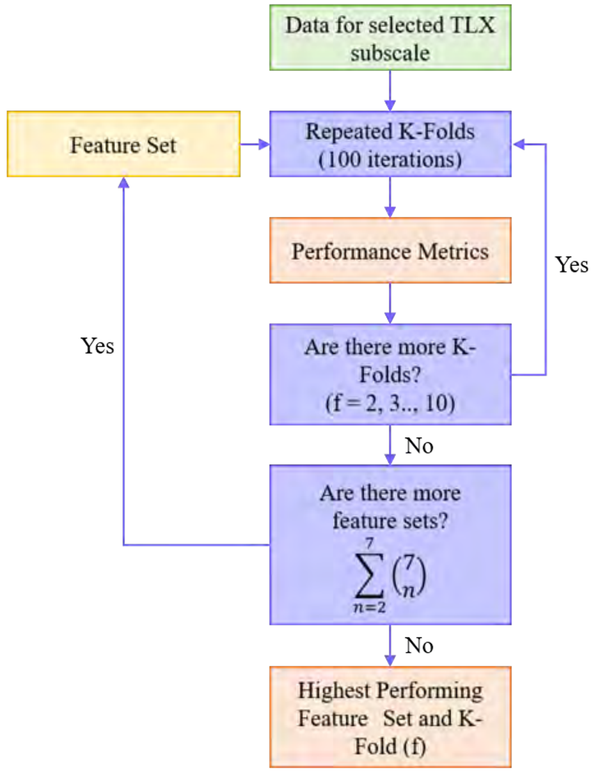

Fig. 4 Systematic approach to building the SVM models for each workload subscale. Blue blocks indicate steps from the K-fold cross-validation process while orange indicate steps that assess model performance metrics.

assigned 1, and 2 vehicle scenarios were assigned 0. The Physical and Performance subscales were scored similarly across subjects and did not provide enough variation for classification. Table 1 provides the binning thresholds for the remaining four subscales.

TLX Classes	Median
Mental	2
Effort	2
Temporal	2
Frustration	1

Table 1 TLX subscale medians used for binning.

B. Features

The metrics captured during each trial were utilized as features and are defined as follows for each vehicle affected by a UVR:

- Number of Uploads: the total number of uploads; an upload updates any combination of altitude, latitude/longitude, route, and speed for a given vehicle
- Initial Response Time (s): time from alert to when the TO finishes the initial vocalization to edit the route
- First Upload (s): time from the start to when the TO finishes the vocalization of the upload for the first time
- Final Upload (s): time from the start to when the TO finishes the vocalization to upload for the final time
- UVR Resolution Time (s): time in which the pilot finishes the vocalization to the time they upload a new route that resolves the UVR conflict
- Service Time (s): total time pilot spends to complete the route modification (including terrain deconfliction)
- Number of Vehicles Affected: total number of vehicles affected by each UVR within a trial

Each feature was binned following the method used for the classes aside from Number of Vehicles Affected (Table 2). The Number of Vehicles Affected was classified by vehicle number (2 or 4) and assigned a numerical value of 1 or 0, respectively.

Features	Median
Number of Uploads	1
Initial Response Time	12
First Upload	30.5
Final Upload	77
UVR Resolution Time	76.5
Service Time	60.5
Number of Vehicles Affected	4 Vehicles = 1 2 Vehicles = 0

Table 2 Feature medians used for binning.

V. Results

Fig. 5 shows the accuracy, precision, recall, and *F*-measure according to number of K-Folds for the highest performing feature set (by accuracy) for the following NASA TLX subscales – Mental Demand, Effort, Temporal Demand, and Frustration Level. Table 3 indicates the features that yielded the highest accuracy for each subscale.

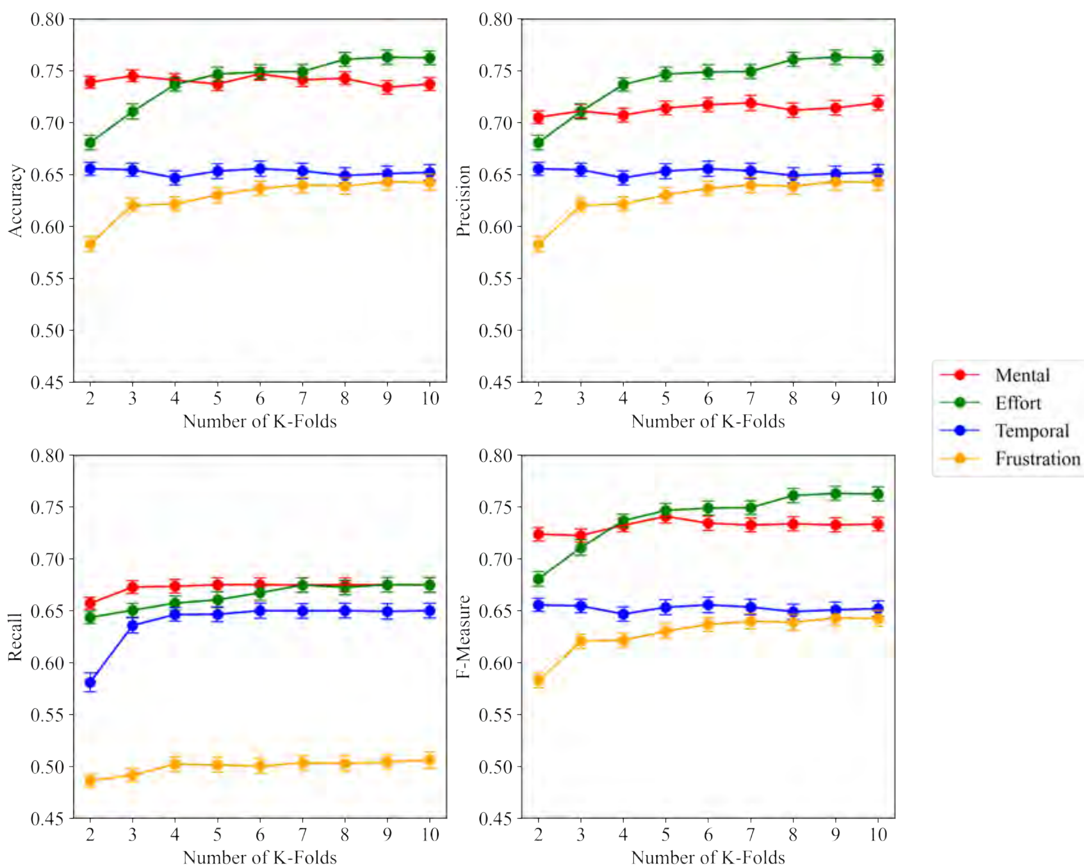

Fig. 5 Performance metrics for the SVM prediction models for Mental Demand, Effort, Temporal Demand, and Frustration Level subscales.

TLX Subscale	Features
Mental	Number of Uploads Initial Response Time First Upload
Effort	Number of Vehicles Affected First Upload
Temporal	Number of Uploads First Upload
Frustration	Number of Uploads Initial Response Time First Upload

Table 3 Input features for each NASA-TLX subscale that produced the best prediction model.

For the Mental demand subscale, the highest performing feature set (Number of Uploads, Initial Response Time, and First Upload) yielded a 74.7% accuracy for 6 K-Folds. Effort (Number of Vehicles Affected and First Upload) scored an accuracy of 76.3% with 9 K-Folds. Temporal Demand (Number of Uploads and First Upload) scored an accuracy of 65.6% with 6 K-Folds and Frustration Level (Number of Uploads, Initial Response Time, and First Upload) scored an accuracy of 64.3% with 9 K-Folds. Notably, there was a <2% and <1% improvement in accuracy between 6 and 9 K-Folds for Effort and Frustration Level, respectively. This indicates that the additional accuracy may not be worth the trade-off for additional computation time.

VI. Discussion

The prediction models (Table 3) show that the method produces fairly accurate models starting about ~60% accuracy up to ~75% accuracy. The highest accuracy was seen in the Effort prediction model followed by Mental Demand, Temporal Demand, and Frustration Level. K-fold values after which very little variation was seen were $f = 8$, $f = 6$, $f = 2$ and $f = 7$ respectively.

Results show that First Upload was consistently a key feature in classifying the NASA TLX subscales with Number of Uploads in 3 of the 4 final feature sets. The time to complete the first upload includes the time to focus their attention and therefore implies that the amount of time needed to reorient their attention has an impact on their overall workload, which is consistent with the previous work from Cummings and Mitchell [14]. Number of Uploads and First Upload were also fairly consistent features. This indicates that UVR and vehicle interaction complexity has a fairly large impact on workload.

The various prediction models have two distinct utilities. The first is that the identified feature combinations that produce the highest accuracy model for each workload subscale can be used as independent variables in future study design to ensure desired workload levels are reached across various conditions. The second use case would be to identify the best feature combination for a given prediction model that would establish a theoretical workload measure. This measure would inform the design of future GCS interfaces and automation technologies while minimizing critical drivers of workload in remote operators. However, future work is needed to identify design features that impact workload.

VII. Conclusion and Future Work

The work in this paper presented a method for developing workload prediction models using a SVM model structure that leveraged K-Fold cross-validation to identify impactful input features. The SVM-based prediction models developed using the proposed methodology was able to generate models for four NASA TLX subscale workload measures: Mental, Effort, Temporal, and Frustration. Results show that the developed prediction models were between ~60-75% accurate with K-fold values between $f = 2$ to $f = 8$. The prediction models have applicability to future study design, as well as, GCS interface design.

Future work will explore prediction models for surrogate workload and will help identify and evaluate factors that may contribute to the overall workload across both the surrogate and subject in this unique teaming paradigm. Additionally, further work could help elucidate how workload is distributed across the two agents. Efforts will also focus on utilizing the workload prediction models in real-time to assess if/when workload level have reached a critical point. The identification of these critically high workload moments will enable the development of automated decision aids that will provide operators with suggestions to temporarily hand off control of some portion of their managed vehicles to other available Tactical Operators until their workload reduces. In doing so, Human-Autonomy Teaming principles will be facilitated and leveraged such that workload is shifted dynamically between team members (i.e., operators) as needed, and overall mission performance will improve.

Acknowledgments

The authors would like to thank Garrett Sadler, Casey Smith, Jillian Keeler, and Jay Shively of the Human Systems Integration Division at NASA Ames Research Center for their work on the development of the design and carrying out data collection for the study whose data was utilized for developing the prediction models presented in this paper. Also, thanks to Dominic Wong and Jonathan Luk of ASRC Federal Data Solutions, LLC for the software development support on the GCS used during the study.

References

[1] Chandarana, M., Hughes, D., Lewis, M., Sycara, K., and Scherer, S., "Planning and Monitoring Multi-Job Type Swarm Search and Service Missions," *Journal of Intelligent & Robotic Systems*, Vol. 101, No. 3, 2021, pp. 1–14.

[2] Liu, Y., Ficocelli, M., and Nejat, G., "A supervisory control method for multi-robot task allocation in urban search and rescue," *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, IEEE, 2015, pp. 1–6.

[3] Hui, X., Bian, J., Yu, Y., Zhao, X., and Tan, M., "A novel autonomous navigation approach for UAV power line inspection," *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, 2017, pp. 634–639.

[4] Zhou, G., Yuan, J., Yen, I.-L., and Bastani, F., "Robust real-time UAV based power line detection and tracking," *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 744–748.

[5] Grzybowski, J., Latos, K., and Czyba, R., "Low-cost autonomous UAV-based solutions to package delivery logistics," *Advanced, Contemporary Control*, Springer, 2020, pp. 500–507.

[6] Magsino, E. R., Say, M. F., and Tan, J. A., "Achieving complete UAV delivery in the presence of motor failures," *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, IEEE, 2020, pp. 1–5.

[7] Chen, J. Y., and Barnes, M. J., "Supervisory control of multiple robots: Effects of imperfect automation and individual differences," *Human Factors*, Vol. 54, No. 2, 2012, pp. 157–174.

[8] Sadler, G., Chandarana, M., Rorie, R. C., Tyson, T. L., Keeler, J. N., Smith, C. L., Shyr, M. C., Wong, D., Scheff, S., and Dolgov, I., "A Remote, Human-in-the-loop Evaluation of a Multiple-Drone Delivery Operation," *AIAA Aviation 2022 Forum*, 2022.

[9] Federal Aviation Administration, "Concept of Operations v2.0, Unmanned Aircraft Systems (UAS) Traffic Management (UTM)," 2020. URL `https://www.faa.gov/sites/faa.gov/files/2022-08/UTM_ConOps_v2.pdf`.

[10] Smith, C. L., Sadler, G., Tyson, T., Brandt, S., Rorie, R. C., Keeler, J., Monk, K., Viramontes, J., and Dolgov, I., "A Cognitive Walkthrough of Multiple Drone Delivery Operations," *AIAA Aviation 2021 Forum*, 2021, p. 2330.

[11] Sheridan, T. B., and Telerobotics, A., *Human supervisory control*, Cambridge, MA: MIT Press, 1992.

[12] Sheridan, T. B., "Supervisory control of remote manipulators, vehicles and dynamic processes: Experiments in command and display aiding," Tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE MAN-MACHINE SYSTEMS LAB, 1983.

[13] Ferrell, W. R., and Sheridan, T. B., "Supervisory control of remote manipulation," *IEEE spectrum*, Vol. 4, No. 10, 1967, pp. 81–88.

[14] Cummings, M. L., and Mitchell, P. J., "Predicting controller capacity in supervisory control of multiple UAVs," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 38, No. 2, 2008, pp. 451–460.

[15] Endsley, M. R., "Toward a theory of situation awareness in dynamic systems," *Situational awareness*, Routledge, 2017, pp. 9–42.

[16] Sheridan, T., "Human supervisory control of robot systems," *Proceedings. 1986 IEEE International Conference on Robotics and Automation*, Vol. 3, IEEE, 1986, pp. 808–812.

[17] Sim, L., Cummings, M. L., and Smith, C. A., "Past, present and future implications of human supervisory control in space missions," *Acta Astronautica*, Vol. 62, No. 10-11, 2008, pp. 648–655.

[18] Chandarana, M., Lewis, M., Sycara, K., and Scherer, S., "Determining effective swarm sizes for multi-job type missions," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 4848–4853.

[19] Kasahara, M., and Cai, K., "Online multi-agent supervisory control for warehouse automation: prioritized tasks," *Proc. 17th International Conference on Automation Science and Engineering*, 2021.

[20] Tatsumoto, Y., Shiraishi, M., Cai, K., and Lin, Z., "Application of online supervisory control of discrete-event systems to multi-robot warehouse automation," *Control Engineering Practice*, Vol. 81, 2018, pp. 97–104.

[21] Lopes, Y. K., Trenkwalder, S. M., Leal, A. B., Dodd, T. J., and Groß, R., "Supervisory control theory applied to swarm robotics," *Swarm Intelligence*, Vol. 10, No. 1, 2016, pp. 65–97.

[22] Breslow, L. A., Gartenberg, D., McCurry, J. M., and Trafton, J. G., "Dynamic operator overload: A model for predicting workload during supervisory control," *IEEE Transactions on Human-Machine Systems*, Vol. 44, No. 1, 2014, pp. 30–40.

[23] Donmez, B., Nehme, C., and Cummings, M. L., "Modeling workload impact in multiple unmanned vehicle supervisory control," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 40, No. 6, 2010, pp. 1180–1190.

[24] Cummings, M. L., Nehme, C. E., Crandall, J., and Mitchell, P., "Predicting operator capacity for supervisory control of multiple UAVs," *Innovations in Intelligent Machines-1*, Springer, 2007, pp. 11–37.

[25] Hamilton, D. B., and Bierbaum, C. R., "Task analysis/workload (TAWL): A methodology for predicting operator workload," *Proceedings of the Human Factors Society Annual Meeting*, Vol. 34, SAGE Publications Sage CA: Los Angeles, CA, 1990, pp. 1117–1121.

[26] Hart, S. G., and Staveland, L. E., "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in psychology*, Vol. 52, 1988, pp. 139–183.

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, No. 85, 2011, pp. 2825–2830. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.