Using Context to Optimize a Functional State Estimation Engine in Unmanned Aircraft System Operations

Kevin Durkee[1(✉)], Scott Pappada[1], Andres Ortiz[1], John Feeney[1], and Scott Galster[2]

[1] Aptima, Inc, Fairborn, USA
{kdurkee, spappada, aortiz, jfeeney}@aptima.com
[2] Air Force Research Laboratory, Dayton, USA
scott.galster@us.af.mil

Abstract. As UAS operations continue to expand, the ability to monitor real-time cognitive states of human operators would be a valuable asset. Although great strides have been made toward this capability using physiological signals, the inherent noisiness of these data hinders its readiness for operational deployment. We theorize the addition of contextual data alongside physiological signals could improve the accuracy of cognitive state classifiers. In this paper, we review a cognitive workload model development effort conducted in a simulated UAS task environment at the Air Force Research Laboratory (AFRL). Real-time workload model classifiers were trained using three levels of physiological data inputs both with and without context added. Following the evaluation of each classifier using four model evaluation metrics, we conclude that by adding contextual data to physiological-based models, we improved the ability to reliably measure real-time cognitive workload in our UAS operations test case.

Keywords: Context · Human performance · Modeling and simulation · Physiological measurement · Workload · UAS

1 Introduction

Unmanned Aircraft Systems (UAS) have grown to become a central capability of the modern United States Air Force by providing essential mission support while preserving the safety of its pilots. Although UAS operations physically remove the human from the aircraft, the human operators of these systems remain an essential component for achieving mission success. As the volume and complexity of UAS operations continues to expand, warfighters will become increasingly vulnerable to undesired cognitive states, such as high workload, stress, fatigue, and vigilance decrements. The ability to monitor these states throughout a mission would be a valuable asset to modern Air Force systems. Measuring cognitive states in relation to task and mission performance would provide the requisite data to detect if, and when, a warfighter has met his/her limits while diagnosing what intervention is best suited to sustaining good performance and obtaining the desired outcomes. By introducing this capability,

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 24–35, 2015.
DOI: 10.1007/978-3-319-20816-9_3

assessments of UAS operators would become integral system parameters about the mission to be proactively monitored and addressed before potential problems occur [1].

There has been a great deal of research on model-based classification techniques to provide real-time operator state monitoring capabilities. Physiological signals have been relied upon as a prominent source of data under the assumption that changes in cognitive activity produce a predictable associated response in physiology. Physiological data are also compelling given their potential to be available at all times and in any work domain, particularly with the emergence and growing affordability of wearable sensors. The majority of research has employed some combination of electroencephalography (EEG) [2], electrocardiogram (ECG) [3], pupillometry [4], or galvanic skin response (GSR) [5]. DARPA's Augmented Cognition (AugCog) was one of the first large-scale efforts to bring this research into warfighter applications [6]. In the Air Force Multi Attribute Task Battery (AF_MATB) environment, Wilson and Russell [7] introduced a novel application of artificial neural networks (ANNs) trained to each individual human performer for real-time mental workload classification using six channels of brain electrical activity, as well as eye, heart, and respiratory signals [7]. Wang et al. (2012) also employed the AF_MATB to introduce a novel hierarchical Bayesian technique that showed promise for cross-subject workload classification [8].

In spite of the many advancements in this line of research, the potential benefits remain inhibited by the inherent noisiness of physiological data. The term "noisy data" refers not only to the robustness of the raw signal itself, but also to the inconsistent and often ambiguous patterns in the processed signals and derived data features. This inconsistency can occur not only across humans, but also within the same individual person over time. Many previous operator state classification efforts have been forced to cater to these caveats and limitations in a variety of ways. For instance, in order to achieve 88 % classification accuracy, Wilson and Russell (2003) focused on highly discrete classifications of "low" versus "high" mental workload, and trained unique ANN models to each individual person [7]. For Wang et al. (2012) although a cross-subject workload classification method was introduced with approximately 80 % classification accuracy across low, medium, and high workload conditions, all eight of their participants' data appeared in both the model training and model testing data sets [8]. Furthermore, the model evaluation metrics reported across much of the published literature tend to focus on how well a *single* output matches that of the intended condition as a whole, rather than evaluating classifier outputs at specific points in time.

While prior research has provided the necessary stepping stones toward a deployable solution, there remains significant room for further innovation. Our recent work has sought to fill several key gaps in physiological-based cognitive state assessment through the development of the Functional State Estimation Engine (FuSE^2), a system designed to derive real-time cognitive state measurements that update frequently (e.g., second-by-second) and provide high-granularity measurement (e.g., 0–100 scale). The FuSE^2 system was initially developed and tested in the AF_MATB [9], and more recently has been applied within a realistic UAS simulation, the Vigilant Spirit Control Station (VSCS) [10]. Although FuSE^2 is capable of on-line supervised learning to adapt to an individual for improving model accuracy, we restrict the scope of this paper solely to cross-subject workload classification since a universal "plug and play" model that does not require per-subject training would be an ideal technological milestone.

The prospect of a cross-subject, physiological-only cognitive state classifier is very challenging, particularly one that can obtain high-resolution measurement. To some extent, there is no guarantee that subtle changes in a person's cognitive state will consistently be reflected in their physiological signals, and any patterns that exist are likely to vary across individuals. This begs the question of what other data sources could be available to help a model classifier "sift through the noise" of physiological signals and ultimately produce more precise measurements. One approach that has not been extensively studied in the operator state modeling literature is the addition of "contextual data" as model inputs alongside a human's real-time physiological signals. Context can generally be defined as any information that can be used to better characterize the situation of an entity [11]. In computer science literature, there is ample evidence that the utilization of contextual data can be used to greatly improve how system applications behave for a variety of purposes [12]. A similar approach could be explored to further optimize physiological-based operator state classification systems. A drawback of this approach is that a particular model classifier could become closely tied to the specific task environment on which it was trained. However, this may be an acceptable tradeoff for system developers since it could reduce a system's dependency on training separate model classifiers for each individual human performer. In addition, by selecting contextual data that are available in a variety of domains – such as system interactions and task performance measures – this maintains the possibility that a cross-person classifier would transfer across some (though assuredly not all) work domains.

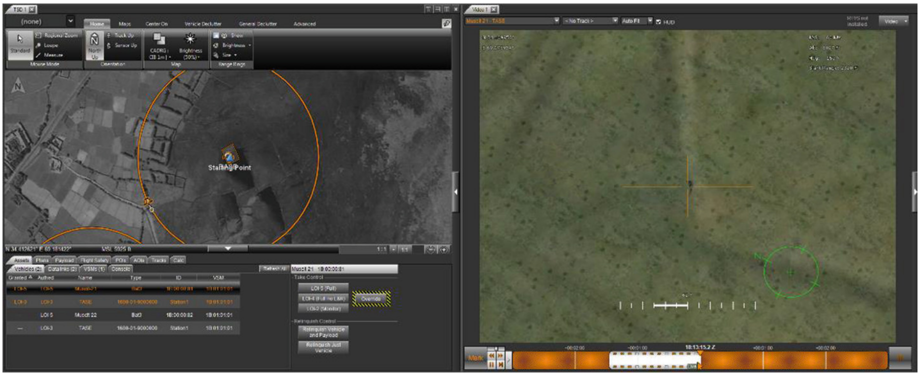
The objective of this work was to explore to what extent contextual data inputs can increase the accuracy of physiological-based cognitive state classifiers in a UAS task environment. In the following sections we review a UAS study that was used to produce data for building real-time workload classifiers within the FuSE^2 system, followed by an analysis with four model evaluation metrics of both second-by-second and aggregated model outputs. We specifically opted to examine the effects of adding two contextual data inputs – human computer interaction (HCI) rate and primary task performance – since these measures were hypothesized to have a predictive relationship with cognitive workload in the UAS study environment, and would be obtainable data inputs in a variety of other task environments.

2 Methods

2.1 Data Collection

Data were collected within a simulated UAS task environment at the Human Universal Measurement and Assessment Network (HUMAN) Laboratory located at Wright-Patterson Air Force Base. We focused exclusively on cognitive workload for this study and the ensuing model development effort so as to constrain the problem space to a single human functional state that has wide applicability, particularly to UAS operations, and a large body of literature to draw from as needed. The UAS task simulation employed the VSCS operator interface (Fig. 1) paired with a Multi-Modal

Communication (MMC) tool for issuing communication requests [13] and a custom-built lights and gauges monitoring display. The primary task objective was to track a high value target (HVT) while keeping the HVT continuously positioned on the center of the UAS sensor crosshairs. Simultaneously, participants conducted two secondary tasks: (1) monitor the lights/gauges display and acknowledge each system event via button presses; and (2) verbally respond to each communication request via the MMC tool. Task difficulty was manipulated by modifying the HVT speed and motion complexity, the number of communication requests, and the number of light/gauge events in each 5 min trial. This task paradigm allowed for a gradual titration of task difficulty across 15 five-minute conditions ranging from easy to hard, which was intended to induce variations in workload and performance for each participant.



Granted	Authed	Name	Type	ID	VSM
LOI-5	LOI-5	Muscit-21	Bat3	[illegible]	[illegible]
LOI-3	LOI-3	TASE	1680-01-9000000	Station1	18:01:01:01
	LOI 5	Muscit 22	Bat3	18:00:00:82	18:01:01:01
—	LOI-3	TASE	1680-01-9000000	Station1	18:01:01:01

Fig. 1. The VSCS operator interface

There were 25 participants with each person completing one training session and one data collection session each. Dependent measures were threefold: (1) a suite of physiological metrics collected during each task condition consisting of six-channel EEG, ECG, off-body eye tracking, respiratory activity, electrodermal activity, and voice analysis features; (2) self-reported NASA Task Load Index (TLX) responses collected at the end of each trial [14]; and (3) system-based performance measures derived from Aptima, Inc.'s Performance Measurement Engine (PM Engine^{TM}) that utilized behavioral and situational data to estimate continuous performance for all three task requirements. NASA TLX responses and condition difficulties yielded a correlation of $r = 0.75$ across all subjects and $r = 0.89$ mean correlation within subjects, suggesting the manipulations were successful at inducing the intended variance in workload.

2.2 Model Development

Using the data collected from this study, a set of model-based classifiers was developed within the FuSE^2 system using machine learning techniques that train each classifier to output second-by-second workload estimates on a 0–100 scale. Adhering to the approach in Durkee et al. [9], we first applied a noise injection algorithm to all NASA TLX responses under the assumption that workload does not remain perfectly static over time. This algorithm derives an estimate of "ground truth" on a second-by-second basis to which the model classifiers are subsequently trained. Because it is impractical to obtain operator responses at very frequent intervals, this algorithm relies on a theoretically-grounded correlate of workload as the basis for injecting this noise. We refer to each series of ground truth estimates as the "desired model output" given each model classifier's attempt to find the best fit based on its feature inputs. A comprehensive training set was then prepared containing all selected feature inputs and the desired model outputs. The training set included data from 19 of the 25 study participants, while the other six participants were randomly selected for model evaluation. A training process was initiated to derive model weights for each classifier based on minimizing error between the feature inputs and the desired model outputs. Three levels of physiological inputs were selected: (1) a "reduced" model consisting of three EEG channels (Fz, Pz, O2) and ECG; (2) a "standard" model consisting of six EEG channels (Fz, Pz, O2, F7, F8, T3) and ECG; and (3) an "expanded" model consisting of the same physiological inputs as the standard model, but with pupillometry included. For each level of physiological input, one classifier was trained with contextual data included and another classifier was trained without contextual data included. A seventh model was also trained consisting solely of contextual data inputs (i.e., no physiological inputs).

2.3 Model Evaluation

After completing the model training process, the next objective was to produce test results in order to evaluate the accuracy of each workload classifier, particularly to assess how the addition of contextual data impacted model accuracy. Workload classifier results were produced through a batch playback of data collected from the six participants excluded from the training set. All six test participants completed the same 15 five-minute trials used to train the model classifiers, thus totaling 90 trials used for evaluation. The batch playback process simulated the production of real-time classifier results by outputting one workload estimate per second on a 0–100 scale for each of the seven models, totaling 300 values per model within each trial. Model accuracy was analyzed via summary statistics in two general ways: (1) second-by-second classifier results (see Sect. 3.1); and (2) aggregated classifier results averaged over entire 5 min trials (see Sect. 3.2). Two model evaluation metrics (similarity score and relative classification accuracy) were used to assess the degree to which each classifier accurately replicated the desired model outputs on a second-by-second basis. In contrast, two other metrics (correlation and absolute difference between average model output and NASA TLX) were used to assess how closely mean classifier output for each trial resembled its respective NASA TLX rating. These four metrics – described further in Sect. 3 – were chosen to balance analyses across the micro- and macro-levels, and to evaluate both patterns and absolute differences compared to ground truth estimates. A secondary objective was to assess the degree to which model accuracy changed as the volume of physiological data inputs is manipulated. A graphical plot is provided for each of the four model evaluation metrics along with discussion of observable trends. Each figure includes results for the three physiological-based model configurations (reduced, standard, and expanded inputs) both with and without contextual data inputs. Results for the context-only model using the specified evaluation metric are also provided as a basis for comparison. Error bars in each graphical plot are shown to illustrate the standard error derived from averaging each metric across the 90 trials.

3 Results and Discussion

3.1 Second-by-Second Workload Classifier Results

The first model evaluation metric is a nonlinear similarity measure known as the correntropy coefficient [15]. For this analysis, we refer to this metric as a "similarity score" in which a higher score indicates a higher degree of pattern similarity between a particular model's actual outputs and the desired model outputs. For the 90 test trials, a correntropy coefficient was derived and each coefficient was adapted to a 0–100 scale to represent the similarity score. A zero value indicates there is no degree of similarity between actual and desired model output, while 100 indicates the actual and desired outputs are identical. We consider the similarity score to be a particularly important metric because it directly gauges how closely each model's second-by-second classifier outputs and trending information matches that of our desired model outputs. This metric's strength is that it focuses squarely on the *shape* of the trend line for each trial's second-by-second model output, while not taking into account whether each individual state classification is at the desired level in an absolute sense. This supplements our other second-by-second model evaluation metric, relative classification accuracy.

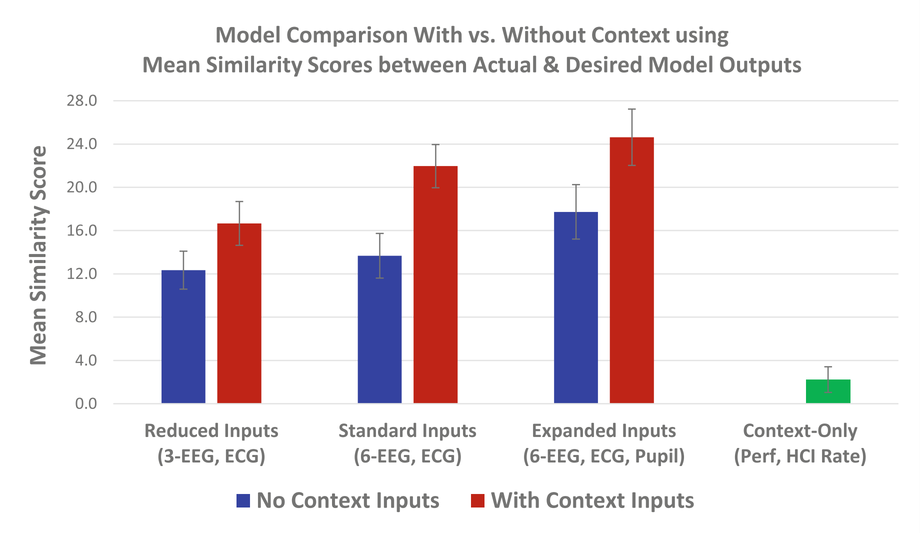


Fig. 2. Comparison of with-context and without-context results for each model configuration using mean similarity scores between 6 test participants' actual & desired model outputs.

Figure 2 illustrates the mean similarity scores for each model as averaged across the 90 test trials. The six physiological-based models yielded two primary trends. First, the additional of contextual data inputs increased the mean similarity score for all three levels of physiological model inputs; the reduced input model increased by 35 %, the standard input increased by 61 %, and the expanded input model increased by 39 %. By comparison, immediately it can be seen this metric highlights a key limitation with the context-only model configuration. With a mean similarity score of only 2.2, the contextual data features by themselves were not able to closely produce the desired model output trend. This is a particularly notable finding given that a context-only model achieved a very low similarity score, yet the addition of contextual data to physiological data produced similarity scores considerably higher than either model could achieve alone. Secondly, as a more general finding, the mean similarity scores increased as the number of physiological inputs increased, resulting in the expanded input model with context producing the highest mean similarity score (24.6). To some extent this pattern is not surprising since this metric most directly reflects the machine learning processes employed during model training. Thus, by increasing the number of feature inputs, the chances of finding reliable model weights will typically increase.

The next metric is classification accuracy of each model's second-by-second outputs relative to desired model output. This metric is intended to supplement the similarity score by examining how frequently the desired *amount* of workload is successfully being measured at any point in time. For this metric we first identified four discrete categories of workload – low, medium-low, medium-high, and high – the boundaries of which were derived from the distribution of all NASA TLX responses. Each individual classifier result was then compared to desired model output at each corresponding point in time for all seven models. Values were flagged as either correct or incorrect based on its "relative" classification state, in which the actual value occurred within one category away from the desired workload category. Though relative classification lacks granularity, it accounts for the possibility of slight misalignment in time between the actual and desired outputs. Additionally, this metric allows for a fairly direct comparison to many prior published approaches that classify "high versus low" workload [7, 8], which is useful as a basis for further model evaluation against alternate approaches.

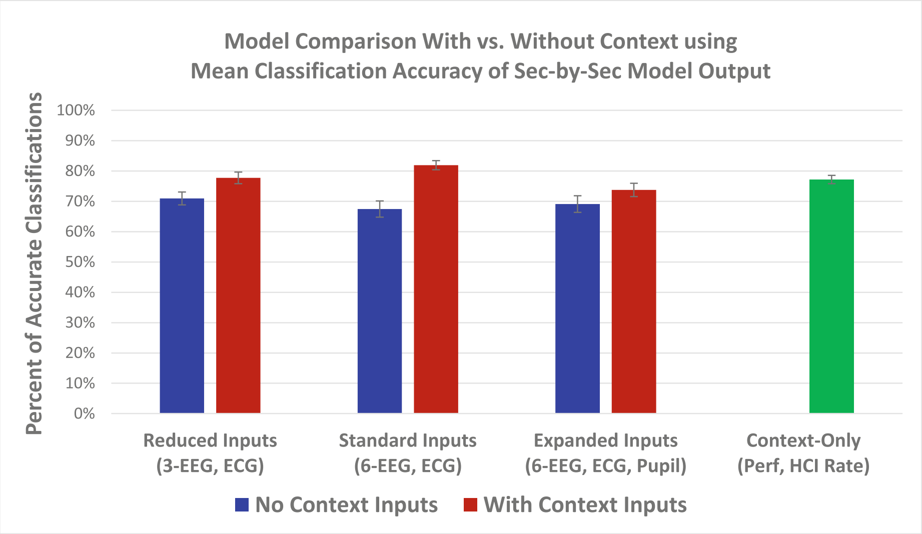


Fig. 3. Comparison of with-context and without-context results for each model configuration using mean classification accuracy of 6 test participants' second-by-second model output.

Figure 3 illustrates the mean classification accuracy of each model's second-by-second outputs across all 90 test trials. Overall, there is low variability across the seven models with a range from 67 % to 82 %. However, there is a consistent trend that by adding context to each of the three physiological-based model configurations, classification accuracy improved in all three cases. The most notable improvement occurred for the standard input model, which produced the lowest classification accuracy without context, yet produced the highest classification accuracy with context added. Furthermore, by adding context to the reduced input and standard input models, this provided another example (much like similarity score) of combining physiological and contextual inputs to provide more accurate classifications than either set of inputs could alone.

3.2 Aggregated Workload Classifier Results

The remaining two model evaluation metrics are: (1) correlation between average model output and NASA TLX; and (2) absolute difference between average model output and NASA TLX. For the correlation analysis, we believed it would be most suitable to derive a Pearson's correlation coefficient (*r*) on a per-person basis to better reflect how a given model tends to track any given person's cognitive workload across the entire length of each trial. As such, the correlation coefficients for each of the six individual test participants and for all seven model classifiers are illustrated in Fig. 4.

Perhaps the most noticeable finding in Fig. 4 is the high correlations achieved by

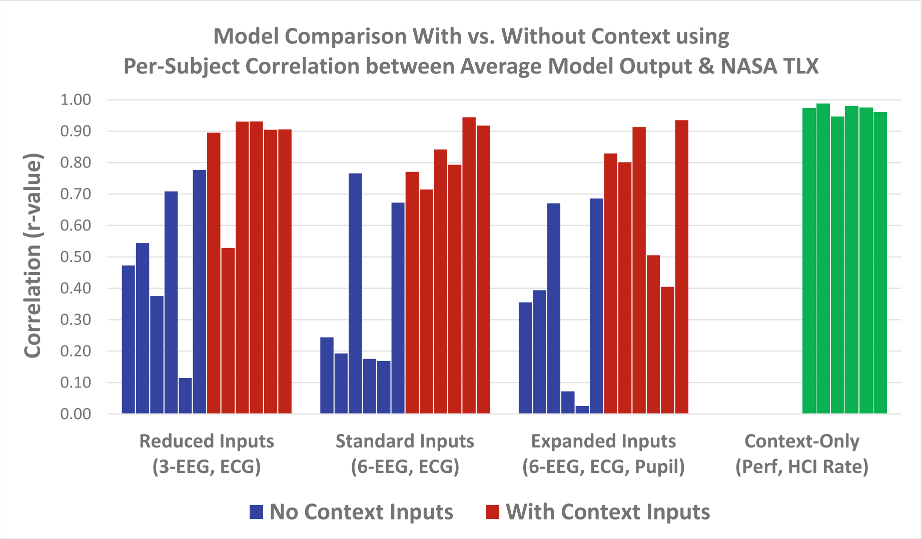


Fig. 4. Comparison of with-context and without-context results for each model configuration using per-subject correlation of NASA TLX and 6 test participants' average model output.

the context-only model ($r = 0.95$–0.99). Although a strong relationship was anticipated between performance, HCI rate, and NASA TLX responses given the nature of the task environment, this result was more pronounced than expected. Another key finding is the degree to which adding context to the three physiological-based models consistently improves each correlation. While these correlations are not quite as high as the context-only model, half of the correlation coefficients are boosted to approximately $r = 0.90$ or higher with the addition of context, and only three of the 18 coefficients remain below $r = 0.70$. Considering the physiological-based models also achieved considerably higher mean similarity scores than the context-only model (see Sect. 3.1), as well as comparable relative classification accuracy, this suggests the blend of physiological and contextual data features may provide the ideal real-time workload assessment capability within this UAS test case. Lastly, one final observation from Fig. 4 is that the correlation coefficients tend to decrease overall as the number of physiological inputs increases. At face value, this finding appears to contrast the observation from mean similarity scores in which a larger number of physiological features resulted in a higher similarity score. This may simply be random chance due to the small number of test participants, or could imply an undesired effect caused by the additional physiological inputs that is not observable in the other metrics. Further evaluation work is needed to gain deeper insight into this observation.

The final model evaluation metric is the absolute difference between average model output and NASA TLX for each trial. This metric provides insight into each model's ability to produce workload classifications that accurately reflect the overall workload induced over the course of an entire trial.

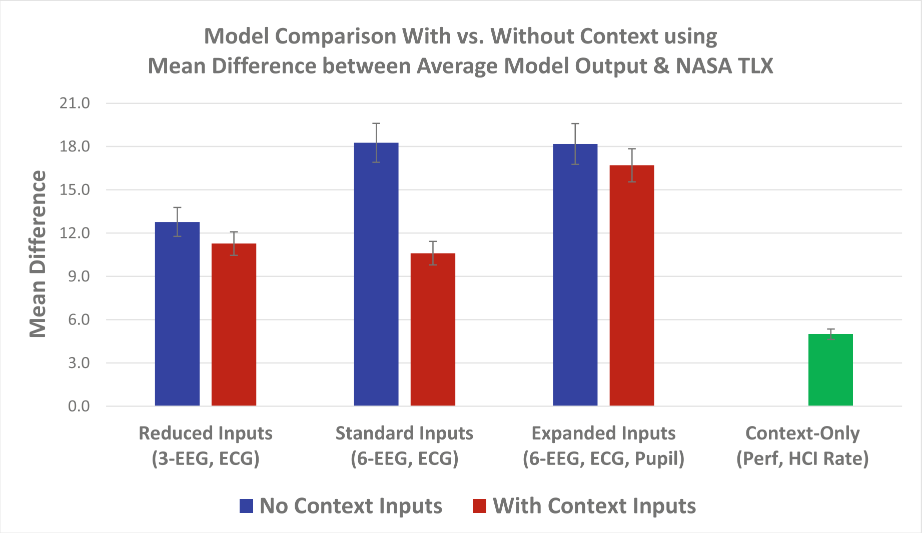


Fig. 5. Comparison of with-context and without-context results for each model configuration using mean difference between NASA TLX and 6 test participants' average model output.

As shown in Fig. 5, the overall trend of these data largely mirrors the correlation results. The context-only model produced the smallest mean difference (5.0) between average model output and NASA TLX rating. Likewise, the addition of contextual data inputs reduced the mean difference for all three levels of physiological-based models (reduced, standard, and expanded inputs). Out of the physiological-based models, the standard input model with context produced the smallest mean difference (10.6) between average model output and NASA TLX ratings, followed closely by the reduced input model (11.3). The standard input model was boosted the most through the additional of contextual data inputs by reducing the mean difference by nearly half (from 18.3 to 10.6). The reduced input model (from 12.8 to 11.3) and expanded input model (from 18.2 to 16.7) were only marginally boosted through the additional of contextual data, though the reduced input model already had a notably smaller difference than the other two models before adding context. Similar to the correlation results in Fig. 4, the model with the largest number of inputs produced the largest mean difference.

4 Conclusions

In summary, we conclude that contextual data collected from our UAS task environment generally enhanced the FuSE^2 system's ability to accurately classify workload for 90 new trials across six test participants. This trend is observed across all four model evaluation metrics and remains consistent, albeit with differing levels of impact, across all three levels of physiological data inputs: 3-channel EEG and ECG ("reduced input model"), 6-channel EEG and ECG ("standard input model"), and 6-channel EEG, ECG, and pupillometry ("expanded input model"). Thorough consideration was given based on the varying perspectives provided by the four model evaluation metrics, in which model results were analyzed on both a second-by-second basis and aggregated basis, as well as analyzing both the absolute differences and patterns of state changes relative to desired model outputs. These results serve as an example of how carefully selected context pertaining to a human performer's behaviors and the mission environment can support a model classifier's ability to translate physiological signals into meaningful assessments of cognitive state.

From a skeptical viewpoint, some of these results could be attributed to either the increase of data inputs in general, or perhaps even to the contextual inputs being the "best" features that account for most of the variance in workload by themselves. While the scope of our test case and analyses prevents ruling out these possibilities completely, there is evidence suggesting this is not necessarily the case. First, it should be noted that the context-only model and reduced input model with context have a small number of inputs compared to the standard input and expanded input models, yet both rated highly on three of the four metrics. It can also be observed that for several of the evaluation metrics, particularly mean difference between average model input and NASA TLX ratings, the expanded input model actually rated lower than the reduced input and standard input models. This may reflect the simple fact that each model depends less on the volume of inputs, and more on having data inputs that provide a discoverable and generalizable indicator of state changes.

It is also necessary to cover an important and perhaps obvious question stemming from these results: is it better to rely simply on a context-only model consisting solely of performance and HCI inputs, while omitting physiological data altogether? This question bears consideration since the context-only model outperformed the six physiological-based models when comparing aggregated model output and NASA TLX ratings (i.e., correlation and absolute difference). At face value, this may challenge the value of physiological data acquisition and convince system developers to rely exclusively on task-specific behavioral data, particularly due to the added cost of introducing physiological sensors. However, there are several counterarguments to this assertion. First, while the aggregated context-only model output was highly predictive of NASA TLX ratings, these model assessments are limited to entire 5 min trials and thus not available on a second-to-second basis. This could be problematic since it may imply the context-only model would not be capable of detecting a sudden and drastic shift in workload with sufficient time to intervene. The similarity scores suggest the physiological-based models with context would be more likely to detect such an occurrence. Second, it bears mentioning that in most cases, contextual data such as HCI rates and primary task performance are likely to be very "task-specific", whereas physiological signals are always present and have some degree of similarity regardless of a person's task requirements. Hence, the accuracy of a context-only model could be more likely to degrade if it were used to assess human operators in new task environments that a trained model has not been exposed to before. Third, it is important to note that as cognitive state monitoring capabilities are developed and tested in different work domains, the utility of different types of contextual data as model inputs is likely to vary substantially. An extensive feature selection process must be accomplished to identify which task-specific indicators exist within a work domain, and the degree of utility they provide to optimize physiological-based models for assessment.

Acknowledgement. Distribution A: Approved for public release; distribution unlimited. This material is based on work supported by AFRL under Contract FA8650-11-C-6236. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL.

References

1. Blackhurst, J., Gresham, J., Stone, M.: The quantified warrior: how DoD should lead human performance augmentation. Armed Forces Journal (2012)
2. Gevins, A., Smith, M.E., Leong, H., McEvoy, L., Whitfield, S., Du, R., Rush, G.: Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. Hum. Factors **40**, 79–91 (1998)
3. Hoover, A., Singh, A., Fishel-Brown, S., Muth, E.: Real-time detection of workload changes using heart rate variability. Biomed. Signal Process. Control **7**, 333–341 (2012)
4. Just, M.A., Carpenter, P.A.: The intensity dimension of thought: pupillometric indices of sentence processing. Can. J. Exp. Psychol. **47**(2), 310–339 (1993)
5. Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G.: Discriminating stress from cognitive load using a wearable EDA device. Technology **14**(2), 410–417 (2010)
6. John, M., Kobus, D.A., Morrison, J.G., Schmorrow, D.: Overview of the DARPA augmented cognition technical integration experiment. Int. J. Hum. Comput. Interact. **17**(2), 131–149 (2004)
7. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum. Factors **45**(4), 635–644 (2003)
8. Wang, Z., Hope, R.M., Wang, Z., Ji, Q., Gray, W.: Cross-subject workload classification with a hierarchical bayes model. NeuroImage **59**(1), 64–69 (2012)
9. Durkee, K., Geyer, A., Pappada, S., Ortiz, A., Galster, S.: Real-time workload assessment as a foundation for human performance augmentation. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2013. LNCS, vol. 8027, pp. 279–288. Springer, Heidelberg (2013)
10. Rowe, A.J., Liggett, K.K., Davis, J.E.: Vigilant spirit control station: a research testbed for multi-UAS supervisory control interfaces. In: Proceedings of the 15th International Symposium on Aviation Psychology, Dayton, OH (2009)
11. Dey, A.: Understanding and using context. Pers. Ubiquitous Comput. **5**(1), 4–7 (2001)
12. Dey, A., Abowd, G., Salber, D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Hum. Comput. Interact. **16**(2), 97–166 (2001)
13. Finomore, V., Popik, D., Dallman, R., Stewart, J., Satterfield, K., Castle, C.: Demonstration of a network-centric communication management suite: multi-modal communication. In: Proceedings of the 55th Human Factors and Ergonomics Society Annual Meeting. HFES, Las Vegas (2011)
14. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Peter, A.H., Najmedin, M. (eds.) Advances in Psychology, vol. 52, pp. 139–183. Elseiver, North-Holland (2006)
15. Xu, J., Bakardjian, H., Cichocki, A., Principe, J.: A new nonlinear similarity measure for multichannel signals. Neural Netw. **21**, 222–231 (2008)