MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220191

Predicting Objective Performance Using Perceived Cognitive Workload Data in Healthcare Professionals: A Machine Learning Study

Karthik Adapa [a,b], Malvika Pillai[b], Shiva Das[a], Prithima Mosaly[c], Lukasz Mazur[a,b]

[a] *Department of Radiation Oncology, School of Medicine, UNC-Chapel Hill, NC, USA*
[b] *Carolina Health Informatics Program, UNC-Chapel Hill, NC, USA*
[c] *Ben Allegretti Consulting, Inc., Stafford, Virginia, USA*

Abstract

Cognitive Workload (CWL) is a fundamental concept in predicting healthcare professionals' (HCPs) objective performance. The study aims to compare the accuracy of the classical model (utilizes all six dimensions of the National Aeronautics and Space Administration Task Load Index (NASA-TLX)) and novel models (utilize four or five dimensions of NASA-TLX) in predicting HCPs' objective performance. We use a dataset from our previous human factors research studies and apply a broad selection of supervised machine learning classification techniques to develop data-driven computational models and predict objective performance. The study findings confirm that classical models are better predictors of objective performance than novel models. This has practical implications for research in health informatics, human factors and ergonomics, and human-computer interaction in healthcare. Findings, although promising, cannot be generalized as they are based on a small dataset. Future studies may investigate additional subjective and physiological measures of CWL to predict HCPs' objective performance.

Keywords:

Machine learning, task performance, cognitive ergonomics

Introduction

Health care is an intrinsically complex field and the increasing complexity and difficulty of tasks impose varying levels of cognitive workload (CWL) which affects healthcare professionals' (HCPs) performance. Suboptimal CWL of HCPs has consistently been associated with inferior task performance and a higher likelihood of errors [12; 18]. Thus, in theory, CWL is a fundamental concept in predicting HCPs' objective performance.

The National Aeronautics and Space Administration Task Load Index (NASA-TLX) is the most widely used subjective measure of CWL of individuals operating in high-risk and time-sensitive industries [8]. The NASA-TLX is a six-item scale that was initially developed to measure perceived CWL in laboratory-based aviation settings and has since been applied to CWL measurement in other domains such as nuclear energy, transportation, and increasingly in health care [8; 9; 13]. The original instrument requires a participant to first perform 15 separate pair-wise comparisons between 6 dimensions (mental, physical and temporal demands, frustration, effort, and performance) and then mark a workload score between low (0) and high (100) for each dimension. Thus, the classical approach involves assessing the composite NASA-TLX score by multiplying the dimension weight with the corresponding NASA-TLX dimension score, summing across all *six* dimensions, and dividing by the number of comparisons (15). However, recent studies in healthcare suggest that using *four* NASA-TLX items (mental, physical, and temporal demands and effort) is a more direct measure of overall CWL than using six dimensions [7; 19]. Thus, there is wide heterogeneity in using NASA-TLX to measure HCPs' CWL which in turn affects the prediction of HCPs' objective performance.

CWL is a multidimensional and complex construct that is often affected by several non-linear factors. Theory-driven and deductive approaches have been utilized in the past to aggregate these factors to define overall CWL and build robust models for predicting objective performance. In recent times, inductive data-driven methodologies such as supervised machine learning (ML) classifiers have been applied to predict objective performance from perceived CWL data. There is uncertainty and ambiguity associated with *how* the non-linear factors affect the overall CWL. This allows the classifiers to learn from data and predict HCPs performance, thus generating alternative inferences. This ML study aims to optimize the prediction of HCPs' objective performance using an optimal subset of NASA-TLX. Previous studies using non-healthcare data suggest that data-driven inductive models generally outperform deductive theory-driven models [17]. However, to the best of our knowledge, no previous study has utilized this methodology to predict objective performance of HCPs using perceived CWL data.

Methods

Dataset

The dataset for this study has been drawn from our previously published studies of HCPs (Table 1). The dataset contains 318 instances from 84 HCPs performing tasks of varying difficulty, context and requiring different human modalities for processing information in a simulated environment. The participants after each task filled out the NASA-TLX questionnaire and objective performance was measured at the end of the task. The objective performance was categorized for each study into low (<=Q1), moderate (>Q1 & <=Q2), and high (>= Q3) based on quartiles. A detailed description of the tasks, self-reporting measures, and objective performance measures can be found in [11], [14], and [16].

Table 1: Dataset for this study

Author	*Grant #*	*Type of Participants*	*Number of participants*	*Objective performance*
Mazur [11]	R18HS025597	Radiation therapists	32	Assessment of procedural compliance and error detection
Mazur [14]	R21HS024062	Resident physicians and fellows	38	Percentage of appropriately managed abnormal test results
Mosaly [16]	R03HS025258	Radiation therapists	14	Assessment of time-out compliance, error detection, and procedural compliance

Machine learning CWL models

We use the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, one of the most widely-used analytics models to implement the ML CWL models [3]. This process involves six stages: identifying goals, understanding data, preparing data, model training, model testing, and deploying the model.

Goals: The study aims to compare classical and novel CWL models in predicting HCPs' objective performance by evaluating classifiers on accuracies and Kappa statistics. The classical CWL model utilizes all the six dimensions of NASA-TLX while the novel CWL models utilize four or five dimensions of NASA-TLX.

Data understanding: The data involved in the creation of classical and novel CWL models include different combinations of NASA-TLX dimensions. Data exploration is the first part of CRISP-DM in which an analytic base table is built for discovering the nature of data and investigating its characteristics (Table 2).

Table 2- Characteristics of features and the outcome (R =Range, C= Categorical)

Features	**Type**	**Miss**	**N**	**Mean**	**SD**
Mental Demand	R	0	314	46.16	24.43
Physical Demand	R	0	314	16.32	14.83
Temporal Demand	R	0	314	28.39	21.84
Effort	R	0	314	39.20	24.33
Performance	R	0	314	32.28	25.96
Frustration	R	0	314	29.33	25.80

Outcome	**Type**	**Miss**	**N**	**Classes (number of instances)**
Objective performance	C	0	314	Low (165) Moderate (74) High (75)

Data preparation: The main aim of this stage is to construct the final dataset for subsequent modeling. Here the dataset is divided into two segments- features and the outcome variable (objective performance) for the following CWL models:

a) Classical model (all 6 dimensions of NASA-TLX as features)
b) Novel model-I (4 dimensions- Mental, Physical, Temporal Demands and Effort)
c) Novel model-II (5 dimensions - Mental, Physical, Temporal Demands, Effort, and Frustration)
d) Novel model-III (5 dimensions- Mental, Physical, Temporal Demands, Effort, and Performance)

Target class imbalance: In ML, the class imbalance impacts the creation of robust models as it tends to favor predicting the majority class over the minority class. In the current dataset, there are two minority classes – medium (74) and high (75) and one majority class –low (165). To solve this issue, we use synthetic minority oversampling techniques (SMOTE)[4], a widely used oversampling method to balance the target classes. Studies suggest that there are more than 85 SMOTE variants[10], but for this study, we implement three widely used SMOTE variants: equal resampling [4], minority resampling [4], and density-based (DB) SMOTE [2]. We use DB-SMOTE because of its ability to avoid model overfitting [6]. We also compare the accuracy of the ML classifiers with and without SMOTE variants. Table 3 shows the number of instances in each objective performance class as well as the total number of instances in the original dataset and the training dataset with and without SMOTE variants.

Table 3- Distribution of objective performance classes in original and training datasets (with/without SMOTE variants)

Dataset	**Objective performance classes**			**Total # of instances**
	Low	*Moderate*	*High*	
Original dataset	165	74	75	314
Training dataset without SMOTE	130	60	61	251
Equal sampling SMOTE	130	130	130	390
Minority Resampling SMOTE	140	100	100	340
DB-SMOTE	130	130	59	319

Model training: The aim of this stage to develop computational models by learning from data. We surveyed the main classifiers and selected the following five ML classification techniques to tackle the CWL modeling problem from different perspectives:

- Similarity-based: K-nearest Neighbors (KNN)
- Information based: Random Forest (RFC)
- Error-based: Nu Support Vector Machine (Nu-SVM)
- Probability-based: Bernoulli Naïve Bayes (BNB)

To develop robust predictive models with a higher degree of generalizability, cross-validation can be used for model training. A random stratified split was conducted on the original dataset (314 instances) with 80% for training and 20% for testing. Stratification was used to ensure that the class imbalance was retained for training and testing (Fig. 1). The training dataset was then oversampled with multiple SMOTE variants. Subsequently, we used 10-fold cross-validation (CV), a widely used training method for small datasets [1]. In this study, after CV, the final models were tested on the held out test set (20% of instances).

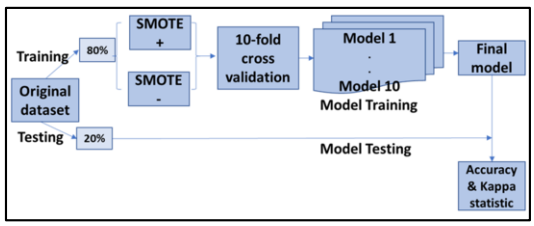

Figure 1. Model training and testing

***Model testing*:** This stage aims at testing the classifiers from the earlier stage, evaluating them on their inferential capacity and accuracy. Overall, 64 final classifiers were built (4 CWL models x 4 classifiers x 4 with and without SMOTE). We selected two metrics- prediction accuracy and the Kappa statistic to evaluate the final classifiers. Accuracy helps in the overall interpretation of a classifier while the Kappa statistic compares observed accuracy with expected accuracy (random chance). The Kappa statistic accounts for random chance and is likely to be less misleading. Further, previous studies show that Kappa statistic is sensitive to imbalanced data and is useful for evaluating multi-class models. Thus, combining Kappa statistic with accuracy provides a more in-depth interpretation than using accuracy alone as an evaluation metric [5].

Results

Cross-validation accuracies

The distribution of accuracies obtained with 10-fold CV for different CWL models (classical model, novel model I-III) is shown in Fig 2. The highest CV accuracy for predicting objective performance is achieved by the classical model followed by the novel models I-III. Further, equal sampling SMOTE achieved higher accuracy than using other SMOTE variants and "without SMOTE" (Fig 3). RFC and Nu-SVM achieved higher accuracy in comparison with other machine learning classifiers as well (Fig 4).

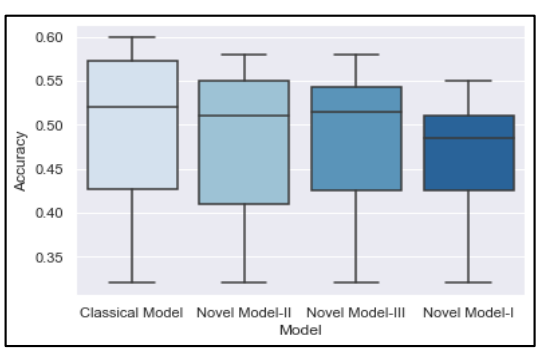

Figure 2.CV accuracies for CWL models

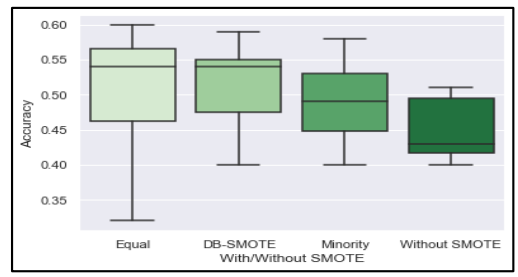

Figure 3.CV accuracies grouped by with and without SMOTE variants

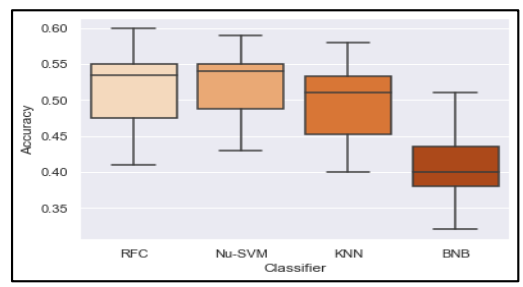

Figure 4. CV accuracies for machine learning classifiers

Testing accuracies and Kappa statistic

The overall average accuracy (0.46, SD 0.11) obtained for 64 classifiers in the model testing phase is lower than that obtained in the model training phase (0.49, SD 0.7). However, the results from the model testing phase show two identical and one non-identical trend in comparison with the model training phase. Figure 5 depicts the distribution of accuracy of different CWL models. As observed in the model testing phase, the classical model has the highest accuracy in predicting objective performance than the novel models. The density plots of testing accuracies also confirm that the classical model with more compact and taller curves are on average superior to the novel models (Fig 6). Further, the distribution of Kappa statistics validates that the classical model is more reliable in predicting objective performance than the novel models (Fig 7). The testing accuracy distribution shows that "without SMOTE" is better than the "SMOTE variants" in building CWL models, which is in contrast to the findings in the model testing phase (Fig 8). Testing accuracy distribution for ML classifiers also shows the same trends as in the model training phase. Fig 9 highlights that Nu-SVM and RFC are the most robust machine learning classifiers to build CWL models for this dataset.

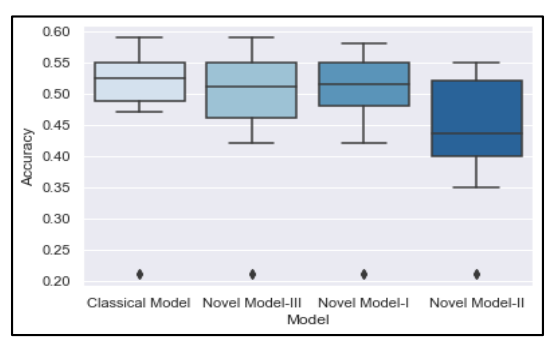

Figure 5. Testing accuracies for CWL models

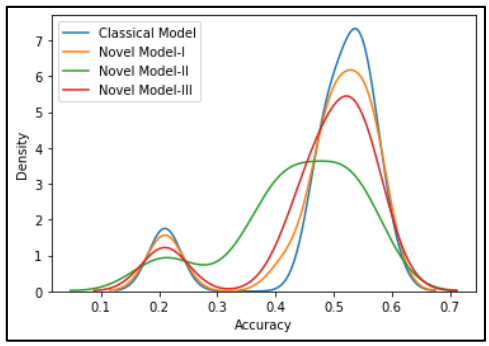

Figure 6. Testing accuracy densities for CWL models

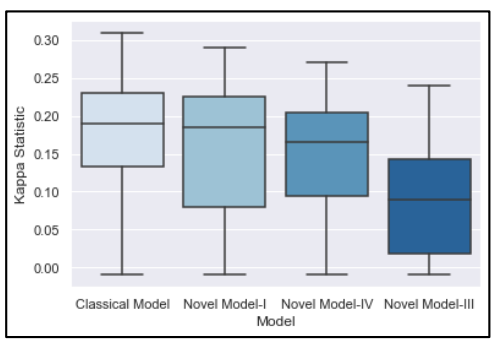

Figure 7. Kappa statistic for CWL models

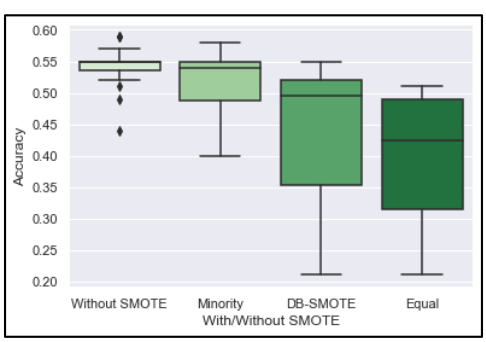

Figure 8. Testing accuracies for with and without SMOTE variants

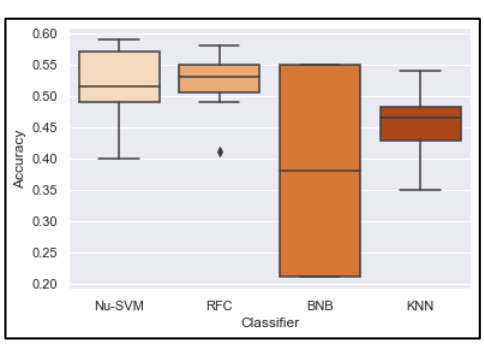

Figure 9 Testing accuracies for machine learning classifiers

Discussion

The findings in this study suggest that the classical model outperforms novel models I-III in predicting HCPs' objective performance based on two evaluation metrics: accuracy and the kappa statistic. We examined 64 final models, used four different supervised ML classifiers, and explored using and not using SMOTE variants to compare the classical and novel CWL models. The classical model consistently outperformed the novel models I-III in CV and testing accuracies and Kappa statistics. Thus, this ML study suggests that the set of 6 dimensions of NASA-TLX (classical model) is a better predictor of HCPs' objective performance in comparison to using 4 or 5 dimensions of NASA-TLX (novel models).

The findings of this study have wide-ranging practical implications for research in the field of health informatics, human factors and ergonomics, and human-computer interaction in healthcare. NASA-TLX dimensions were originally designed to be weighted, but more recently, researchers have started using the raw scores of dimensions which are averaged or added to create an overall CWL estimate [8]. Furthermore in healthcare, using other approaches such as confirmatory factor analysis, the number of NASA-TLX dimensions to calculate overall CWL has been reduced to four [15; 19]. This study utilizes a data-driven ML methodology to suggest that all six dimensions of NASA-TLX should be used to better predict HCPs' objective performance.

We also found that the testing accuracies even for the best-supervised machine learning classifiers (RFC and Nu-SVM) vary from 0.40 to 0.59, indicating that either more data is needed to build better CWL models, or more independent features and other non-linear factors influencing CWL are necessary to increase the accuracies. However, these results are in line with current research on CWL and the widely accepted difficulties in predicting human performance [17].

We also report that the testing accuracies of the "without SMOTE" dataset were higher than "with SMOTE" variants although all the SMOTE variants outperformed "without SMOTE" on CV accuracies. Previous studies suggest that SMOTE greatly improves minority class detection and overall classification performance. However, the limited size of the dataset and overall low accuracies may have enabled "without SMOTE" to classify the majority class correctly while misclassifying minority classes and thus having higher accuracy. Thus, this justifies the argument that accuracy alone is not a reliable predictor for minority classes, and therefore, these findings have to be interpreted cautiously.

Our study results have important implications in quantifying overall CWL of HCPs. Despite recent evidence in favor of utilizing a four-item NASA-TLX as a quick measure of overall CWL of HCPs [15; 19], our study suggests that all dimensions must be measured to assess overall CWL of HCPs.

Comparison with prior work

Overall, our findings are consistent with the Moustafa et al. study that used non-healthcare data (405 total instances) to develop data-driven computational models of CWL to predict objective performance [17]. Important differences between the present study and their study were that they considered two subjective measures of CWL (NASA-TLX and Workload profile) and within NASA-TLX used raw scores of all dimensions of NASA-TLX, weights of all dimensions, and total TLX scores as independent features. However, in this study, we compared raw scores of six dimensions of NASA-TLX with four and five dimensions of NASA-TLX to predict HCPs objective performance. Further, we used multiple SMOTE oversampling techniques on the data, while Moustafa et al. used

only DB-SMOTE for oversampling [17]. To answer the research question/goal of the study, we utilized 64 classifiers while Moustafa et al. utilized 16 classifiers for comparison.

This study has several limitations. Our results are based on a small dataset of HCPs (62% participants are radiation therapists) at a single large academic medical center drawn from three Agency for Healthcare Research and Quality grants. The objective performance in each of the grants is assessed differently and hence findings from this study, although promising, cannot be generalized. Further, we use only raw scores of NASA-TLX dimensions as independent features, which limits the generalizability of the study findings.

Conclusions

This human-centered machine learning study suggests that a set of 6 dimensions of NASA-TLX (classical model) is a better predictor of objective performance of healthcare professionals than using 4 or 5 dimensions of NASA-TLX (novel models). This finding has implications for research in health informatics and human factors in healthcare. Further empirical evidence for this study confirms that testing accuracies for predicting the objective performance of HCPs even with the best ML classifiers vary from 0.4 to 0.59, thus highlighting the need for additional independent features to better predict objective performance. Further empirical research is required using weighted scores of NASA-TLX, total TLX scores, other self-reporting CWL assessment techniques (e.g., Workload profile), physiological measures, data from other HCPs performing primary and secondary tasks in different healthcare contexts (e.g., operation theater, simulation in safety-critical environments, electronic health records) to confirm if building data-driven models of CWL would improve our prediction of HCPs' objective performance.

Acknowledgments

We thank Prof. Arcot Rajasekar, School of Information and Library Science for assisting in the conception of the study.

References

[1] S. Arlot and A. Celisse, A survey of cross-validation procedures for model selection, *Statistics surveys* **4** (2010), 40-79.

[2] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique, *Applied Intelligence* **36** (2012), 664-684.

[3] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth, The CRISP-DM process model. CRIP-DM Consortium 310, in, CRISP-DM consortium, 1999.

[4] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research* **16** (2002), 321-357.

[5] M. Fatourechi, R.K. Ward, S.G. Mason, J. Huggins, A. Schlögl, and G.E. Birch, Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets, in: *2008 Seventh International Conference on Machine Learning and Applications*, IEEE, 2008, pp. 777-782.

[6] C. Gavin and N. Tablot, On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, *Journal of Machine Learning Research* (2010).

[7] E. Harry, C. Sinsky, L.N. Dyrbye, M.S. Makowski, M. Trockel, M. Tutty, L.E. Carlasare, C.P. West, and T.D. Shanafelt, Physician task load and the risk of burnout among US physicians in a national survey, *Joint Commission Journal on Quality and Patient Safety / Joint Commission Resources* **47**, 76-85.

[8] S.G. Hart, Nasa-Task Load Index (NASA-TLX); 20 Years Later, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **50** (2006), 904-908.

[9] C.J. Jacobson, S. Bolon, N. Elder, B. Schroer, G. Matthews, J.P. Szaflarski, M. Raphaelson, and R.D. Horner, Temporal and subjective work demands in office-based patient care: an exploration of the dimensions of physician work intensity, *Medical Care* **49** (2011), 52-58.

[10] G. Kovács, Smote-variants: A python implementation of 85 minority oversampling techniques, *Neurocomputing* **366** (2019), 352-354.

[11] L.M. Mazur, R. Adams, P.R. Mosaly, M.P. Stiegler, J. Nuamah, K. Adapa, B. Chera, and L.B. Marks, Impact of Simulation-Based Training on Radiation Therapists' Workload, Situation Awareness, and Performance, *Advances in radiation oncology* **5** (2020), 1106-1114.

[12] L.M. Mazur, P.R. Mosaly, L.M. Hoyle, E.L. Jones, and L.B. Marks, Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks, *Practical radiation oncology* **3** (2013), e171-177.

[13] L.M. Mazur, P.R. Mosaly, M. Jackson, S.X. Chang, K.D. Burkhardt, R.D. Adams, E.L. Jones, L. Hoyle, J. Xu, J. Rockwell, and L.B. Marks, Quantitative assessment of workload and stressors in clinical radiation oncology, *International Journal of Radiation Oncology, Biology, Physics* **83** (2012), e571-576.

[14] L.M. Mazur, P.R. Mosaly, C. Moore, and L. Marks, Association of the usability of electronic health records with cognitive workload and performance levels among physicians, *JAMA network open* **2** (2019), e191709.

[15] E.R. Melnick, E. Harry, C.A. Sinsky, L.N. Dyrbye, H. Wang, M.T. Trockel, C.P. West, and T. Shanafelt, Perceived electronic health record usability as a predictor of task load and burnout among US physicians: mediation analysis, *Journal of Medical Internet Research* **22** (2020), e23382.

[16] P.R. Mosaly, R. Adams, G. Tracton, J. Dooley, K. Adapa, J.K. Nuamah, L.B. Marks, and L.M. Mazur, Impact of workspace design on radiation therapist technicians' physical stressors, mental workload, situation awareness, and performance, *Practical radiation oncology* (2020).

[17] K. Moustafa, S. Luz, and L. Longo, Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques, in: *Human mental workload: models and applications,* L. Longo and M.C. Leva, eds., Springer International Publishing, Cham, 2017, pp. 30-50.

[18] V.L. Patel and T.G. Kannampallil, Cognitive informatics in biomedicine and healthcare, *Journal of Biomedical Informatics* **53** (2015), 3-14.

[19] H.L. Tubbs-Cooley, C.A. Mara, A.C. Carle, and A.P. Gurses, The NASA Task Load Index as a measure of overall workload among neonatal, paediatric and adult intensive care nurses, *Intensive & critical care nursing : the official journal of the British Association of Critical Care Nurses* **46** (2018), 64-69.

Address for correspondence

Karthik Adapa, MBBS, MPH, Email : karthikk@live.unc.edu