2019

Analysing the Impact of Machine Learning to Model Subjective Mental Workload: A Case Study in Third-Level Education

Karim Moustafa
Technological University Dublin

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Recommended Citation

Moustafa K., Longo L. (2019) Analysing the Impact of Machine Learning to Model Subjective Mental Workload: A Case Study in Third-Level Education. In: Longo L., Leva M. (eds) Human Mental Workload: Models and Applications. H-WORKLOAD 2018. Communications in Computer and Information Science, vol 1012. Springer, Cham. DOI: 10.1007/978-3-030-14273-5_6

Analysing the Impact of Machine Learning to Model Subjective Mental Workload: A Case Study in Third-Level Education

Conference Paper · February 2019

DOI: 10.1007/978-3-030-14273-5_6

2 authors:

Karim Moustafa
Technological University Dublin - City Campus

Luca Longo
Technological University Dublin - City Campus

Analysing the impact of Machine Learning to model subjective Mental Workload: a case study in third-level education

Karim Moustafa, Luca Longo*

[1] School of Computer Science, Technological University Dublin, Dublin, Republic of Ireland
[2] ADAPT, the global centre of excellence for digital content technology, Dublin, Republic of Ireland
*luca.longo@dit.ie

Abstract. Mental workload measurement is a complex multidisciplinary research area that includes both the theoretical and practical development of models. These models are aimed at aggregating those factors, believed to shape mental workload, and their interaction, for the purpose of human performance prediction. In the literature, models are mainly theory-driven: their distinct development has been influenced by the beliefs and intuitions of individual scholars in the disciplines of Psychology and Human Factors. This work presents a novel research that aims at reversing this tendency. Specifically, it employs a selection of learning techniques, borrowed from machine learning, to induce models of mental workload from data, with no theoretical assumption or hypothesis. These models are subsequently compared against two well-known subjective measures of mental workload, namely the NASA Task Load Index and the Workload Profile. Findings show how these data-driven models are convergently valid and can explain overall perception of mental workload with a lower error.

1 Introduction

Assessing human mental workload is fundamental in the disciplines of Human-Computer Interaction and Ergonomics [13,53]. Through mental workload, human performance can be predicted and used for designing interacting technologies and systems aligned to the limitations of the human mental limited capabilities [26]. However, despite its theoretical utility, and after decades of research, it is still an umbrella construct [21,12,30]. In the last 50 years, researchers and scholars have devoted their effort to the design and development of models of mental workload that can act as a proxy for assessing human performance [15,9,47,35]. Mental Workload (MWL) is a complex psychological construct, believed to be multi-dimensional and composed of several factors. Various approaches have been developed to measure and to aggregate these factors into an overall index of mental workload [50,28,22]. The vast majority of these are theory-driven, which means that they utilise theoretical hypothesises and beliefs for assessing MWL deductively. Also, even if theoretically sound, these models are rather ad-hoc and they mainly adopt basic operators for aggregating factors together, with the implicit assumption of their linearity and often additivity. However, it is argued that MWL is far from being a linear phenomenon and the application of non-linear computational approaches can advance its modelling. Additionally, instead of using theoretical knowledge, it is argued that data-driven approaches are likely to offer a significant improvement in the developmend of models of mental workload [56]. In particular, Machine Learning (ML) is one of these approaches that has been recently considered in MWL modelling. For example, researchers have started applying ML techniques using physiological or task performance measures [51,55]. Other studies employing ML have shown promising results as in [40,49,38].

This research study aims at investigating the impact of supervise modelling techniques, hardly borrowed from machine learning, in the creation of models of MWL by employing subjective self-reporting features from humans. In detail, this study compares traditional subjective models of MWL, namely the NASA Task Load Index (NASA-TLX) [14] and the Workload Profile (WP) [50], against data-driven models produced by a number of ML techniques. Concisely, this paper attempts to answer the research question: *Can machine learning techniques help build data-driven models of mental workload that have a better face validity than the Nasa Task Load Index and the Workload Profile?*

The rest of this paper is organised as follows. Section 2 describes related work in the field of MWL measurement, with an emphasis on subjective approaches. It then discusses the gaps in the literature that motivate the need of non-linear modelling methods for mental workload. Section 3 introduces the design of a comparative study and it describes the research methodology adopted for building data-driven models of mental workload. Section 4 presents the findings and critically evaluates them with a rigorous comparison against the selected MWL baseline instruments, namely the NASA-TLX and the Workload Profile. This comparison is performed by computing the convergent and face validity of the induced MWL models from data. Finally, Section 5 concludes the paper by highlighting its contribution and suggesting future work.

2 Related Work

The importance of measuring MWL has arisen from the crucial need of predicting human performance [23,25,26]. In turn, human performance plays a central role in the design of interactive technologies, interfaces as well as educational and instructional material [31,29,23,36,24,37,27]. Measuring mental workload is not a trivial task [48]. Various measures exist, with different advantages and disadvantages, and they can be clustered in three main classes:

- subjective measures - this class refers to the subjective perception of the operator who is executing a specific task or interacting with an underlying system. Subjective measures, also referred to as self-reporting measures, rely on a direct estimation of individual differences such as emotional state, level of stress, the effort devoted to the task and its demand. The perception of users usually can be gathered by means of surveys or questionnaires in the post-task phase [13]. This category includes measures such as the NASA Task Load Index (NASA-TLX) [14], the Workload Profile (WP) [50] based on the Multiple Resource Theory [52], and the Subjective Workload Assessment Technique (SWAT) [42];
- task performance measures - this category includes primary and secondary task measures. These measures focus on quantifying the objective performance of humans in relation to a specific task under execution. Example include the number of errors, the time needed and the resources used to accomplish a task or the reaction time to a secondary task [34,**?**];
- physiological measures - this class relies on the analysis of the physiological responses of a human executing a task. Examples include the heart rate, EEG brain signals, eye movements and skin conductivity [4,35].

Self-reporting subjective measures are based upon the assumption that only the human involved with a task can provide accurate and precise judgements about the experienced mental workload. They are often employed post-task and are easy to be administered. For these reasons, they are appealing to many practitioners and are the focus of this paper. However, they contribute to an overall description of the mental workload experienced on a task with no information about its temporal variation. The category of task performance measures is based upon the belief that the mental workload experienced by an individual becomes relevant only if it impacts system performance. Primary task measures are strongly connected to the concept of performance since they provide objective and quantifiable measures of error or human success. Secondary task measures can be gathered during task execution and are more sensitive to mental workload variation. However, they might influence the execution of the primary task and in turn influence mental workload. The class of physiological measures considers responses of the body gathered from the individual interacting with an underlying task/system. The assumption is that they are highly correlated to mental workload. Their utility lies in the interpretation and analysis of psychological processes and their effect on the state of the body over time, without demanding an explicit response by the human. However, they require specific equipment and trained operators minimising their employability in real-world tasks.

2.1 Subjective Measurements Methods

Two out of the several subjective measures of mental workload developed in the last decades are the NASA Task Load Index (NASA-TLX) [14] and the Workload Profile (WP) [50]. Since these have been selected as baselines in this research study, their detailed description follows. NASA-TLX is a mental workload assessment tool developed by the the National Aeronautics and Space Administration agency. It was originally conceived to assess the mental workload of pilots during aviation tasks. Subsequently, it was adopted in other fields and used as a benchmark in many research studies as for instance in [46,43,44,45,27]. The original questionnaire behind this instrument can be found in [14]. The NASA-TLX scale is built upon six dimensions and an additional pair-wise comparison among these dimensions. This comparison is used to give weights to the six dimensions as shown in equation 1.

$$NASA-TLX_{MWL} = \left(\sum_{i=1}^{6} d_i \times w_i \right) \frac{1}{15} \tag{1}$$

The Workload Profile (WP) is based on the Multiple Resource Theory (MRT) that was introduced by prof. Wickens [52]. The WP index is derived from eight dimensions: perceptual/central processing, response processing, spatial processing, verbal processing, visual processing, auditory processing, manual responses, and speech responses. In WP, the operator is asked to report the proportion of attentional resources elicited during task execution. The final mental workload score is a sum of the eight factors, as shown in equation 2.

$$WP_{MWL} = \sum_{i=1}^{8} d_i \tag{2}$$

For a detailed information about the scales used by the two mental workload instruments described above, the reader is referred to [22].

2.2 Machine Learning and data-driven methods for mental workload modeling

Machine learning (ML) is a subfield of Artificial Intelligence that focuses on creating models from data. It can be seen as a method of data analysis for automated analytical model building. It focuses on automatic procedures than can learn from data and identify patterns with minimal human intervention. ML can be supervised, unsupervised or semi-supervised. On one hand, supervised ML aims to build mathematical models from a set of data that contains both the inputs and the desired output (supervisory data). On the other hand, unsupervised ML takes only input data and it is aimed at finding structures, patterns, and groups or clusters in it. Semi-supervised ML employs both the above learning mechanisms and it occurs when not all the inputs have an associated output. A number of research studies have employed ML for mental workload modeling. For example, [41,16] analysed physiological brain signals, gathered by functional Near-Infrared Spectroscopy (fNIRS), with unsupervised ML. [49] and [40] employed supervised ML respectively using speech data and linguistic/keyboard dynamics of the operators to predict her/his mental workload. [8]

and [32] adopted supervised ML for mental workload assessment using features extracted from eye movements. Similarly, supervised ML was used to predict levels of cognitive load in driving tasks employing physiological eye movements and primary task measures such as braking, acceleration, and steering angles [55]. Recently, the multi-model approach of combining multiple physiological measures for mental workload assessmnt has emerged demonstrating an enhancement over using individual techniques separately [1,20]. Supervised ML has also been employed with subjective self-reporting data [38] and compared against well-known self-reporting measures.

3 Design And Methodology

In order to tackle the research question formalised in section 1, a comparative research study was designed to evaluate the accuracy of data-driven models, built with supervised machine learning versus two subjective baselines models of mental workload, namely the NASA-TLX and the WP, as shown in figure 1. Two criteria for evaluating MWL models have been selected, in line to other studies in the literature [43,22]: convergent [5] and face validity [39]. The definitions of these two forms of validity adopted here are shown in Table 1. Existing data has been used and the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) has been followed for constructing MWL models [7].

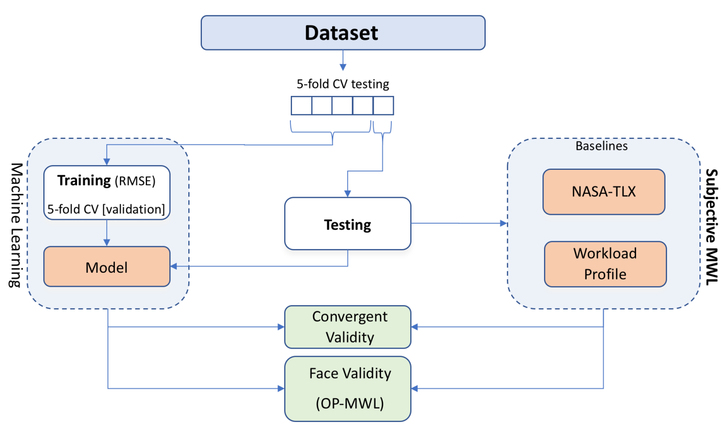


Fig. 1: The design of a comparative study aimed at comparing data-driven models of mental workload, built with supervised machine learning, against two subjective baseline models.

Table 1: Criteria for comparing mental workload models

Name	Description	Statistical Tools
Convergent Validity	It aims to determine whether different MWL assessment measures are theoretically related.	Correlation coefficient of the MWL scores produced my baseline models vs ML models.
Face Validity	It aims to determine the extent to which a measure can actually grasp the construct of MWL.	Error of a MWL model in predicting a self-reported perception of MWL.

3.1 Dataset, context and participants

The dataset selected for this research study has been formed in an educational context. More specifically, recruited participants were students who attended classes of the *Research Design and Proposal Writing* module, in a master course in the School of Computing, at Dublin Institute of Technology. Four different topics have repeatedly been delivered in four consecutive semesters, from 2015 to 2017. ('Science', 'The Scientific Method', 'Planning Research', 'Literature Review'). These topics were delivered adopting three different instructional formats:

1. The first format focused on the transmission of information with a traditional direct-instruction method – from lecturer to students – by projecting slides on a whiteboard and describing them verbally.
2. The second format included the delivery of the same content, as developed using the first format, as multimedia videos, pre-recorded by the same lecturer. Videos were built by employing the principles of the Cognitive Theory of Multimedia Learning [33]. Further details can be found in [27];
3. The third format included a collaborative activity conducted after the delivery of the video, as developed in the second format. The goal of this activity was to improve the social construction of the information through dialogue among students divided in groups.

The number of classes, their length and the number of students are summarised in Table 2. Students were of 16 nationalities (19-54 years; mean=31.7, std=7.5). For each class, students were randomly split into two groups. They respectively received the questionnaire associated to the NASA-TLX and the WP. In addition to this, students were asked to answer an additional question on overall perception of MWL, hereinafter referred to as the *Overall Perception of Mental Workload (OP-MWL)*, on a discrete scale from 0 to 20 (figure 2). Those students who agreed to participate in the experiment received a consent form, approved by the ethics committee of the Dublin Institute of Technology, and a study information sheet. These forms describe the theoretical framework of the study, the confidentiality of the data, and the anonymisation of their personal information. Thus, two sub-datasets were formed, one containing the answers of the NASA-TLX questionnaire, and one related to the answers related to the WP questionnaire, respectively containing 145 and 139 samples.

Table 2: Number of classes for each format, number of students in each class and their length in minutes

Lecture	**Format 1**			**Format 2**			**Format 3**		
	classes	students	mins	classes	students	mins	classes	students	mins
Science	2	14,17	62,60	1	26	18	1	16	60
Scientific Method	1	23	46	2	18,18	28,28	1	18	50
Research Planning	1	20	54	2	22,22	10,10	1	9	79
Literature Review	1	21	55	1	24	19	1	16	77

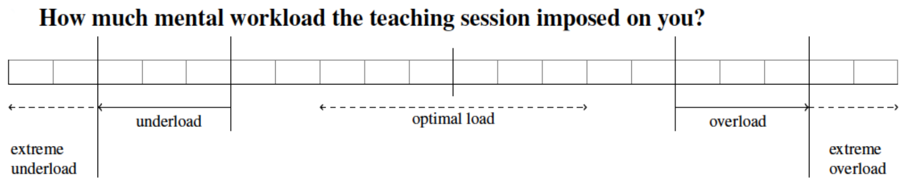


Fig. 2: Scale of the question for measuring the overall perception of mental workload (OP-MWL).

3.2 Machine learning for training mental workload models

Supervised machine learning was employed to train models of mental workload from collected data. The dependent feature is the overall perception of mental workload provided by students (OP-MWL) while the independent features are the questions of the NASA-TLX and the WP instruments.

Data Understanding - Three sets of independent features were formed, as described in the summary table 3. This helped understand the nature of the data and it allowed the investigation of its characteristics, such as the type of features, their values and ranges. The table also shows the normality of the distributions of each feature and its skewness. Figure 3 depicts the distribution of the target variable (the overall perception of mental workload OP-MWL).

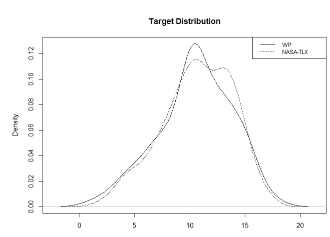


Fig. 3: Distribution of the target variable: the overall perception of mental workload provided by students (OP-MWL).

	Type	n	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
Feature set 1: questions of the NASA-TLX											
Mental	R	145	10.04	3.42	10	1	20	19	-0.04	-0.34	0.28
Physical	R	145	6.31	4.19	6	1	20	19	0.63	-0.22	0.35
Temporal	R	145	9.22	3.41	10	1	20	19	-0.01	0.16	0.28
Performance	R	145	8.72	3.73	9	2	17	15	0.17	-0.92	0.31
Frustration	R	145	7.55	3.93	7	1	19	18	0.43	-0.57	0.33
Effort	R	145	9.89	4.02	10	1	20	19	0.13	-0.18	0.33
Feature set 2: pairwise comparisons of the NASA-TLX											
Temporal_vs_frustration	C	145	1.19	0.4	1	1	2	1	1.54	0.37	0.03
Performance_vs_mental	C	145	1.48	0.5	1	1	2	1	0.07	-2.01	0.04
Mental_vs_physical	C	145	1.09	0.29	1	1	2	1	2.84	6.13	0.02
Frustration_vs_performance	C	145	1.81	0.4	2	1	2	1	-1.54	0.37	0.03
Temporal_vs_effort	C	145	1.62	0.49	2	1	2	1	-0.49	-1.77	0.04
Physical_vs_frustration	C	145	1.45	0.5	1	1	2	1	0.21	-1.97	0.04
Performance_vs_temporal	C	145	1.41	0.49	1	1	2	1	0.38	-1.87	0.04
Mental_vs_effort	C	145	1.38	0.49	1	1	2	1	0.49	-1.77	0.04
Physical_vs_temporal	C	145	1.8	0.4	2	1	2	1	-1.48	0.21	0.03
Frustration_vs_effort	C	145	1.79	0.41	2	1	2	1	-1.43	0.05	0.03
Physical_vs_performance	C	145	1.91	0.29	2	1	2	1	-2.84	6.13	0.02
Temporal_vs_mental	C	145	1.7	0.46	2	1	2	1	-0.85	-1.29	0.04
Effort_vs_physical	C	145	1.08	0.28	1	1	2	1	3	7.03	0.02
Frustration_vs_mental	C	145	1.81	0.39	2	1	2	1	-1.6	0.55	0.03
Performance_vs_effort	C	145	1.43	0.5	1	1	2	1	0.26	-1.94	0.04
Feature set 3: questions of the Workload Profile											
Solving_deciding	R	139	11.17	3.93	11	2	20	18	-0.18	-0.51	0.33
Response_selection	R	139	9.92	4.34	10	1	20	19	-0.16	-0.72	0.37
Task_space	R	139	8.74	4.71	9	1	20	19	0.07	-0.96	0.4
Verbal_material	R	139	12.48	3.8	13	2	20	18	-0.57	-0.32	0.32
Visual_resources	R	139	12.24	3.79	13	3	20	17	-0.45	-0.42	0.32
Auditory_resources	R	139	12.78	3.69	13	4	20	16	-0.3	-0.57	0.31
Manual_response	R	139	9.46	5.05	10	1	20	19	-0.03	-0.92	0.43
Speech_response	R	139	8.82	5.03	9	1	20	19	0.14	-0.98	0.43
Dependent features											
$OP-MWL$ (NASA group)	R	145	10.68	3.19	11	2	17	15	-0.41	-0.39	0.27
$OP-MWL$ (WP group)	R	139	10.47	3.37	10	1	18	17	-0.38	-0.19	0.29

Table 3: Summary Table (ST) of the dataset features and targets (R=Range, C=Categorical)

Data Preparation - The final datasets to be used for training purposes were subsequently constructed. Two datasets were formed:

- *dataset NASA-TLX*: this includes all the NASA-TLX features, in addition to the binary preferences which emerged from the pairwise comparison of the original instrument (Feature sets 1+2 of Table 3).
- *dataset WP*: this includes all the eight features of WP (Feature set 3 of Table 3).

The *dataset NASA-TLX* had 41 missing values spotted in 11 records (all in the pair-wise comparison part) so, due to the limited amount of available data, imputation was performed. The *K-Nearest Neighbours* (KNN) algorithm was applied to estimate missing values based on the concept of similarity. This algorithm has demonstrated good performance without affecting the quality of data [2,17]. K represents the number of nearest instances to be considered while calculating the missing instance.

Data Modelling - This stage is aimed at inducing models of mental workload by learning from available data rather than making ad-hoc theory-driven models. An assumption made is that the aggregation of those factors believed to model mental workload is non-linear. Tackling the complex problem of MWL modelling, and in the spirit of the *No-Free-Lunch* theorem [54] – stating that there is not one best approach that always outperforms the other – different supervised machine learning algorithms for non-linear regression were chosen. Each learning strategy encodes a distinct set of assumptions, that means different inductive biases. Additionally, a linear method based on probability was also selected for comparison purposes:

- Information-based: Random Forest by Randomization regression (Extra Trees: Extremely Randomized Trees) [11];
- Similarity-based: K-Nearest Neighbours regression [18];
- Error-Based: Support Vector Regression (Radial basis function kernel) [3];
- Probability-based: Bayesian Generalised Linear Model regression [10].

The datasets were randomly split into 5 partitions of equal size, non overlapping. Four of these were used for training purposed (80% of the data) and the held-out set for testing purposes (20%) of the data. The process was repeated 5 times, and at each time, the held-out set was different. The parameters employed in each regression technique have been automatically tuned through a random search approach (number of randomly selected predictors and number of random cuts for extra trees, the number of neighbours for KNN and sigma and regularisation term for SVM) Additionally, 5-fold cross validation has been used in each training phase and the Root Mean Square Error as metric (RMSE) for fitting the overall perception of mental workload (OPMWL). Therefore, one is expected to have 5 *surrogate models*, for each training phase. The best one, that means the one with less RSME, was kept as the *final induced model*. Since, the process was repeated 5 times, as per figure 1, one is expected to be left with 5 induced models for each regression technique.

Model Evaluation - In order to evaluate the final induced models from data, the following error metrics are evaluated [19,6]:

- Mean Squared Error (MSE) (eq. 3). It is the most common metric for the evaluation of regression-based models. The higher the value the worse the

model. It is useful if observations contain unexpected value that are important. In case of a single very bad prediction, the squaring will make the error even worse, thus skewing the metric and overestimating the badness of the regression model (range $[0, \infty)$];
- Root Mean Squared Error (RSME) (eq. 4). It is the square root of the MSE and it has the ability to present the variance on the same scale of the target variable. (range $[0, \infty)$; here $[0, 20]$);
- Mean Absolute Error MAE (eq. 5). It is a linear score and all the individual differences between expected and predicted outcome are weighted equally in the average. Contrarily to MSE, it is not that sensitive to outliers. (range $[0, \infty)$];

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2} \tag{4}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\,(y_i - \widehat{y}_i)\,| \tag{5}$$

with y_i is the actual expected value, $\widehat{y}_i$ is the model's prediction

4 Results And Evaluation

4.1 Accuracy of the final induced models

Figure 4 depicts the box-plots containing the RMSE values for training. According to the previous design, each box plot contains 5 points, one for each final induced model trained with 80% of the data. It can be observed that, in most of the cases, the final induced models, trained with the NASA-TLX features (feature sets 1 + 2 of Table 3), have always lower RSME than those models built upon the WP features (feature set 3 of table 3), even if this is not significant. This denotes that the selected regression techniques can train a model similarly and consistently. Also, since it is in the scale [0,20], it denotes the small error in fitting the target feature (OP-MWL). In fact, errors on average, lies between 1 and 5, across the selected regression techniques, with mean around 3. It can be also noted that the mean of the error of the Bayesian generalised linear models is higher than the others, non-linear model, preliminary confirming the previous hypothesis of non-linearity of the independent features. This means that the non-linear models can better learn the non-linear aggregation of the independent features.

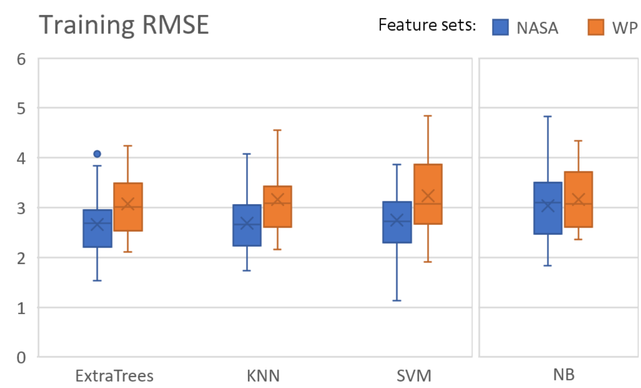


Fig. 4: The distributions of the RSME of the final induced models, grouped by features sets (NASA Task Load Index, Workload Profile). Each bar contains 5 values, one for each model grouped by the regression technique.

4.2 Convergent validity of the induced models

The convergent validity of the induced models is assessed by calculating the Spearman's correlation between their inferred MWL scores, and the scores produced by the baseline models (NASA-TLX, WP) using the testing sets. Figure 5 shows these correlation coefficients in box-plots, each containing 5 values corresponding to the 5 trained models tested with the 5 testing sets of 20% each. The Spearman's correlation statistic was used because the assumptions behind the Pearson's correlation statistics were not met. Generally, a moderate/high positive coefficients have been found (with $p < 0.05$) indicating that the inferences of the induced models, built with machine learning, are valid since they correlate with the baseline models. Also, these results are in line to the recommendation of [5] whereby convergent validities above $\rho = 0.70$ are recommended, whereas those below $\rho = 0.50$ should be avoided.

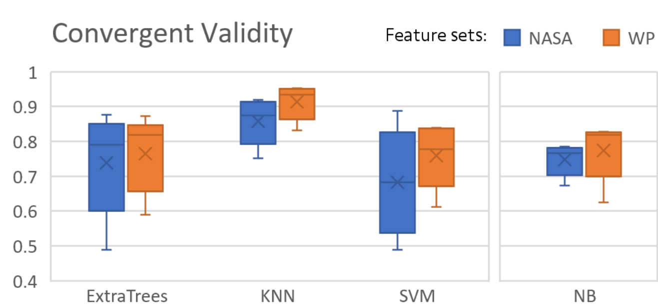
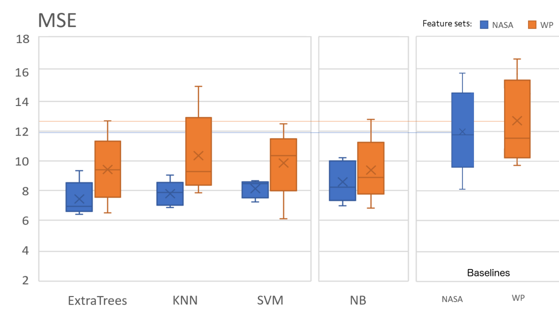


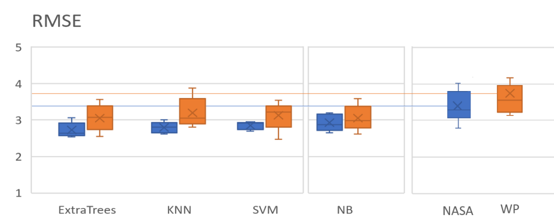
Fig. 5: Convergent validity of the final induced models.

4.3 Face Validity of Induced Models

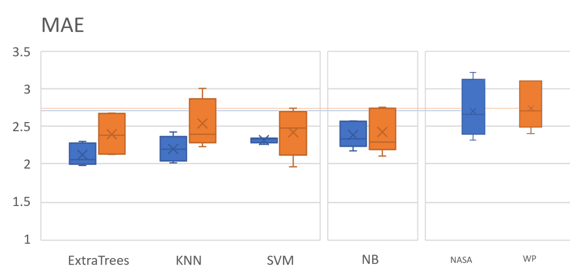
Face Validity was computed to measure the extent to which the final induced models can actually grasp the construct of Mental Workload. This was determined by computing the error of the final induced models, and the selected baselines, in predicting the overall perception of mental workload (OP-MWL) with the testing data, that means they are evaluated with unseen data. Figures 7, 8 and 9 show the scatterplots of this comparison while figure 6 depicts the MSE, RMSE and MAE values. As in the previous case, each box-plot contains 5 values corresponding to the 5 error obtained with the testing sets of 20% each)



(a) Mean Square Errors



(b) Room Mean Square Errors



(c) Mean Absolute Errors

Fig. 6: The distributions of the errors of the final induced models and baseline models, grouped by features set used (NASA-TLX or Workload Profile). Each bar contains 5 points, one for each model grouped by the regression technique.

Firstly, the situation is consistent with the training error: slightly higher for the induced models trained with the WP features. However, the error boundaries for the testing sets are narrower than those achieved during training. In fact, the RSME values, regardless of the regression techniques employed, have mean around 3, with shorter box-plots, suggesting a good degree of generalisability of the induced models. Also it can be seen that the mean of the errors produced by the baseline models is always higher than those produced by the induced models. In other words, the baseline models generate indexes of mental workload that are always more distant to the overall perception of mental workload, reported by subjects, when compared to the distance of the inferences produced by the machine learning models.

4.4 Discussion

Findings are promising and show how subjective mental workload can be modelled with a higher degree of accuracy using data-driven techniques, when compared to traditional subjective techniques, namely the NASA Task Load Index and the Workload Profile, used as baselines. In detail, an analysis of the convergent validity of the data-driven models, learnt from data by employing supervised machine learning regression techniques, against the selected baseline models, show how these are theoretically related. In other words, if we believe that the baseline models actually measure mental workload, so we can do the same with the data-drive models. With this confidence, a subsequent analysis of their face validity showed how data-driven models can approximate the perception of overall mental workload, as reported by subjects, with a higher degree of precision (less error) when compared to the selected baselines. This means that data-driven models covering the concept it purports to measure, that means Mental Workload, with a higher precision. Findings are indeed restricted to the dataset under consideration, but they motivate further research in this space.

5 Conclusion

This work presents an assessment of the ability of machine learning techniques to model mental workload. The motivation behind this work was to shift from state-of-the-art MWL modelling techniques – mainly theory-driven – to automated learning techniques able to induce MWL models from data. Specifically, a number of learning regression techniques have been selected to induce models of mental workload employing features gathered from users subjectively. These features included the answers to the questionnaires of the NASA Task Load Index and the Workload Profile, two baseline mental workload self-reporting measures chosen for comparative purposes. The induced models were compared against the two selected baselines through an assessment of their convergent and face validity. Convergent validity was aimed at determining whether the induced models were theoretically related to the selected baselines, known to model the construct of mental workload. Face validity was aimed at determining whether the induced models could actually cover the concept it purports to measure, that means Mental Workload. The former validity was assessed through a correlation analysis of the mental workload scores produced by the induced models and the selected baselines. The latter validity was assessed by investigating the error of the machine learning models and the baselines to predict an overall perception of mental workload subjectively reported by subjects, after the completion of experimental tasks in third level education.

The findings of this experiment confirm that supervised machine learning algorithms are potential alternatives to traditional theory-driven techniques for modeling mental workload. Machine learning poses itself as a seed for an efficient mechanism that facilitates the understanding of the construct of mental workload, the relationship of its factors and their impact to task performance. A viable direction for future work would be to extend the current experiment with an in depth evaluation of the importance of each feature for predicting the overall perception of mental workload. Subsequently, simpler mental workload models could be created containing the most important features. This can increase the understanding of the complex but fascinating construct of mental workload and contribute towards the ultimate goal of building a highly generalisable model that can be employed across fields, disciplines and experimental contexts.

References

1. Aghajani, H., Garbey, M., Omurtag, A.: Measuring mental workload with eeg+ fnirs. Frontiers in human neuroscience 11, 359 (2017)
2. Batista, G.E., Monard, M.C.: A study of k-nearest neighbour as an imputation method. HIS 87, 251–260,48 (2002)
3. Bennett, K.P., Campbell, C.: Support vector machines. ACM SIGKDD Explorations Newsletter 2(2), 1–13 (2000),
4. Cain, B.: A review of the mental workload literature. Tech. rep., Defence Research and Development Canada Toronto Human System Integration Section; 2007. Report No. : RTO-TRHFM-121-Part-II. Contract No (2004)
5. Carlson, K.D., Herdman, A.O.: Understanding the impact of convergent validity on research results. Organizational Research Methods 15(1), 17–32 (2012)
6. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development 7(3), 1247–1250 (2014)
7. Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., Wirth, R.: The crisp-dm process model. The CRIP–DM Consortium 310 (1999)
8. Cortes Torres, C.C., Sampei, K., Sato, M., Raskar, R., Miki, N.: Workload assessment with eye movement monitoring aided by non-invasive and unobtrusive micro-fabricated optical sensors. In: Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology. pp. 53–54. ACM (2015)
9. Fan, J., Smith, A.P.: The impact of workload and fatigue on performance. In: International Symposium on Human Mental Workload: Models and Applications. pp. 90–105. Springer (2017)

10. Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S.: A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics 2(4), 1360–1383 (2008)
11. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning 63(1), 3–42 (2006)
12. Hancock, P.A.: Whither workload? mapping a path for its future development. In: International Symposium on Human Mental Workload: Models and Applications. pp. 3–17. Springer (2017)
13. Hancock, P.A., Meshkati, N.: Human mental workload. Elsevier (1988)
14. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in Psychology 52(C), 139–183 (1988)
15. Hart, Sandra, G.: NASA-task load index (NASA-TLX); 20 years later. Human Factors and Ergonomics Society Annual Meting pp. 904–908 (2006)
16. Hincks, S.W., Afergan, D., Jacob, R.J.: Using fnirs for real-time cognitive workload assessment. In: International Conference on Augmented Cognition. pp. 198–208. Springer (2016)
17. Jonsson, P., Wohlin, C.: An evaluation of k-nearest neighbour imputation using likert data. In: 10th International Symposium on Software Metrics, 2004. Proceedings. pp. 108–118 (Sept 2004)
18. Kotsiantis, S.B.: Supervised Machine Learning: A Review of Classification Techniques. Informatica 31(2), 249–268 (2007),
19. Kvålseth, T.O.: Cautionary note about r 2. The American Statistician 39(4), 279–285 (1985)
20. Liu, Y., Ayaz, H., Shewokis, P.A.: Multisubject "learning" for mental workload classification using concurrent eeg, fnirs, and physiological measures. Frontiers in human neuroscience 11, 389 (2017)
21. Longo, L.: Formalising Human Mental Workload as a Defeasible Computational Concept. Ph.D. thesis, Trinity College Dublin (2014)
22. Longo, L.: A defeasible reasoning framework for human mental workload representation and assessment. Behaviour and Information Technology 34(8), 758–786 (2015)
23. Longo, L.: Designing medical interactive systems via assessment of human mental workload. In: Int. Symposium on Computer-Based Medical Systems. pp. 364–365 (2015)
24. Longo, L.: Mental workload in medicine: foundations, applications, open problems, challenges and future perspectives. In: Computer-Based Medical Systems (CBMS), 2016 IEEE 29th International Symposium on. pp. 106–111. IEEE (2016)
25. Longo, L.: Subjective usability, mental workload assessments and their impact on objective human performance. In: IFIP Conference on Human-Computer Interaction. pp. 202–223. Springer (2017)
26. Longo, L.: Experienced mental workload, perception of usability, their interaction and impact on task performance. PloS ONE 13(8), 1–36 (08 2018),
27. Longo, L.: On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: A case study in a third-level course. In: Proceedings of the 10th International Conference on Computer Supported Education, CSEDU 2018, Funchal, Madeira, Portugal, March 15-17, 2018, Volume 2. pp. 166–178 (2018),
28. Longo, L., Barrett, S.: Cognitive effort for multi-agent systems. In: International Conference on Brain Informatics. pp. 55–66. Springer (2010)

29. Longo, L., Dondio, P.: On the relationship between perception of usability and subjective mental workload of web interfaces. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. vol. 1, pp. 345–352. IEEE (2015)
30. Longo, L., Leva, M.C.: Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers, vol. 726. Springer (2017)
31. Longo, L., Rusconi, F., Noce, L., Barrett, S.: The importance of human mental workload in web-design. In: 8th International Conference on Web Information Systems and Technologies. pp. 403–409 (April 2012)
32. Mannaru, P., Balasingam, B., Pattipati, K., Sibley, C., Coyne, J.: Cognitive context detection in uas operators using eye-gaze patterns on computer screens. In: Next-Generation Analyst IV. vol. 9851, p. 98510F. International Society for Optics and Photonics (2016)
33. Mayer, R.E.: Cognitive Theory of Multimedia Learning, p. 4371. Cambridge Handbooks in Psychology, Cambridge University Press, 2 edn. (2014)
34. Meshkati, N., Loewenthal, A.: An Eclectic and Critical Review of Four Primary Mental Workload Assessment Methods: A Guide for Developing a Comprehensive Model. Advances in Psychology 52(1978), 251 – 267 (1988),
35. Mijović, P., Milovanović, M., Ković, V., Gligorijević, I., Mijović, B., Mačužić, I.: Neuroergonomics method for measuring the influence of mental workload modulation on cognitive state of manual assembly worker. In: International Symposium on Human Mental Workload: Models and Applications. pp. 213–224. Springer (2017)
36. Mohammadi, M., Mazloumi, A., Kazemi, Z., Zeraati, H.: Evaluation of Mental Workload among ICU Ward's Nurses. Health promotion perspectives 5(4), 280–7 (2015),
37. Monfort, S.S., Sibley, C.M., Coyne, J.T.: Using machine learning and real-time workload assessment in a high-fidelity uav simulation environment. In: Next-Generation Analyst IV. vol. 9851, p. 98510B. International Society for Optics and Photonics (2016)
38. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques. In: Longo, L., Leva, M.C. (eds.) Human Mental Workload: Models and Applications. pp. 30–50. Springer International Publishing, Cham (2017)
39. Nevo, B.: Face validity revisited. Journal of Educational Measurement 22(4), 287–293 (1985)
40. Ott, T., Wu, P., Paullada, A., Mayer, D., Gottlieb, J., Wall, P.: ATHENA A zero-intrusion no contact method for workload detection using linguistics, keyboard dynamics, and computer vision. In: Communications in Computer and Information Science. vol. 617, pp. 226–231 (2016),
41. Pham, T.T., Nguyen, T.D., Van Vo, T.: Sparse fnirs feature estimation via unsupervised learning for mental workload classification. In: Advances in Neural Networks, pp. 283–292. Springer (2016)
42. Reid, G.B., Nygren, T.E.: The subjective workload assessment technique: A scaling procedure for measuring mental workload. In: Advances in psychology, vol. 52, pp. 185–218. Elsevier (1988)
43. Rizzo, L., Dondio, P., Delany, S.J., Longo, L.: Modeling mental workload via rule-based expert system: A comparison with NASA-TLX and workload profile. IFIP Advances in Information and Communication Technology 475, 215–229 (2016),

44. Rizzo, L., Longo, L.: Representing and inferring mental workload via defeasible reasoning: A comparison with the NASA task load index and the workload profile. In: Proceedings of the 1st Workshop on Advances In Argumentation In Artificial Intelligence co-located with XVI International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy, November 16-17, 2017. pp. 126–140 (2017)
45. Rizzo, L., Longo, L.: Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study. In: Proceedings of the 2nd Workshop on Advances In Argumentation In Artificial Intelligence co-located with XVII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018), Trento, Italy, November 20-23, 2018. pp. 11–26 (2018)
46. Rubio, S., Daz, E., Martn, J., Puente, J.M.: Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. Applied Psychology 53(1), 61–86 (2004),
47. Smith, A.P., Smith, H.N.: Workload, fatigue and performance in the rail industry. In: International Symposium on Human Mental Workload: Models and Applications. pp. 251–263. Springer (2017)
48. Smith, K.T.: Observations and issues in the application of cognitive workload modelling for decision making in complex time-critical environments. In: International Symposium on Human Mental Workload: Models and Applications. pp. 77–89. Springer (2017)
49. Su, J., Luz, S.: Predicting cognitive load levels from speech data. Smart Innovation, Systems and Technologies 48, 255–263 (2016)
50. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics 39(3), 358–381 (1996)
51. Walter, C., Cierniak, G., Gerjets, P., Rosenstiel, W., Bogdan, M.: Classifying mental states with machine learning algorithms using alpha activity decline. In: ESANN 2011, 19th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 27-29, 2011, Proceedings (2011),
52. Wickens, C.D.: Multiple resources and mental workload. Human factors 50(3), 449–455 (2008)
53. Wickens, C.D.: Mental workload: assessment, prediction and consequences. In: International Symposium on Human Mental Workload: Models and Applications. pp. 18–29. Springer (2017)
54. Wolpert, D.H.: The supervised learning no-free-lunch theorems. In: Soft computing and industry, pp. 25–42. Springer (2002)
55. Yoshida, Y., Ohwada, H., Mizoguchi, F., Iwasaki, H.: Classifying Cognitive Load and Driving Situation with Machine Learning. International Journal of Machine Learning and Computing 4(3), 210–215 (2014)
56. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics 58(1), 1–17 (2015)

Appendix

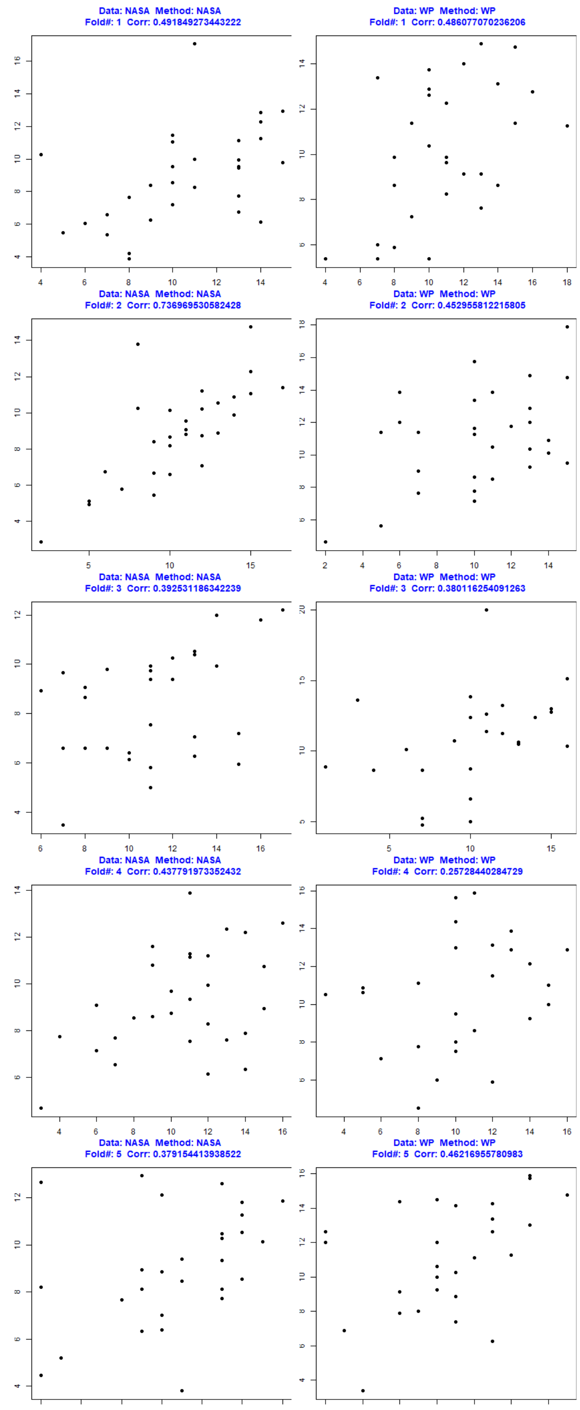


Fig. 7: Scatterplots of the overall perception of mental workload reported by subjects (OP-MWL) (x-axis) and the prediction of the induced models (y-axis) for the NASA-TLX (Left) and the Workload Profile (Right) grouped by fold

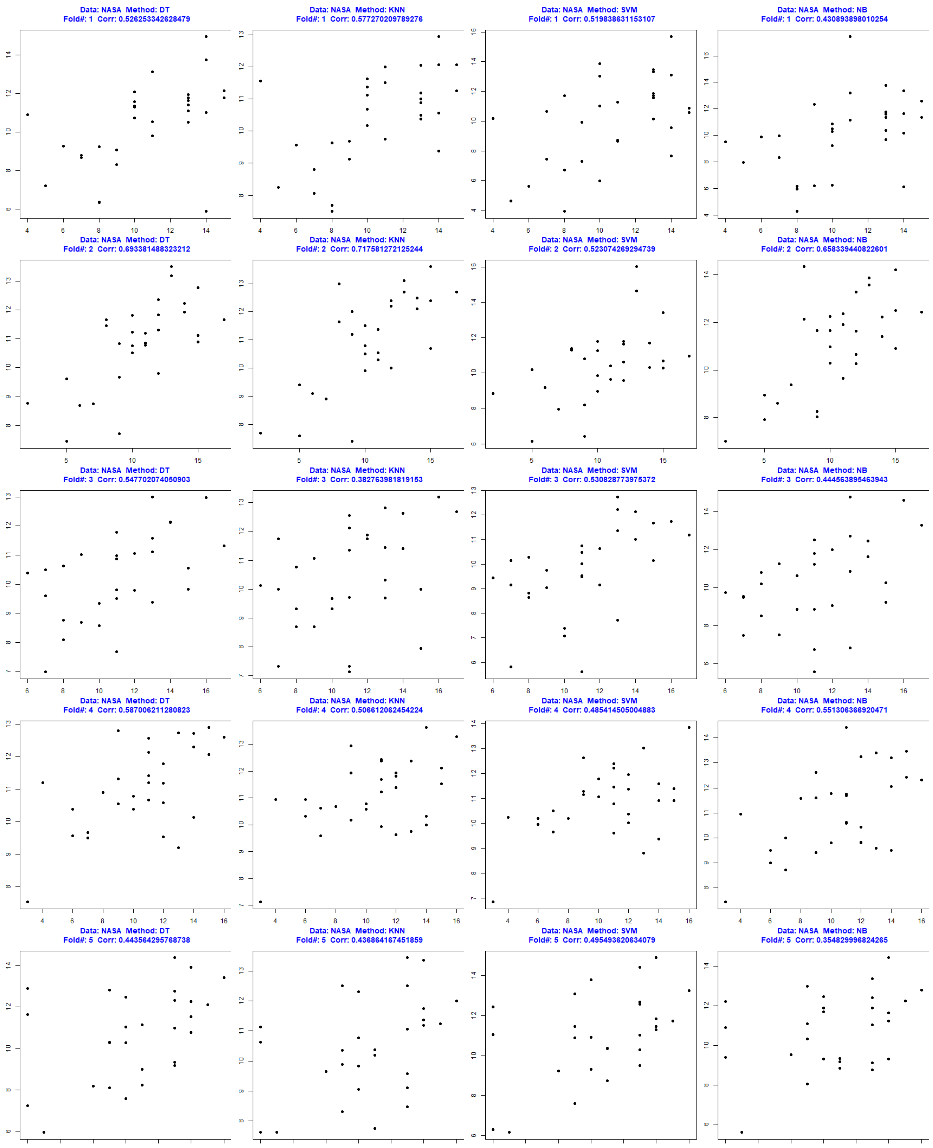


Fig. 8: Scatterplots of the overall perception of mental workload (x-axis), as reported by subjects and the prediction of the induced models (y-axis) for the 5 models produced by the regression algorithms (Extra trees: col 1; KNN: col 2; SVR: col 3; NB: col 4) employing the features of the NASA Task Load Index

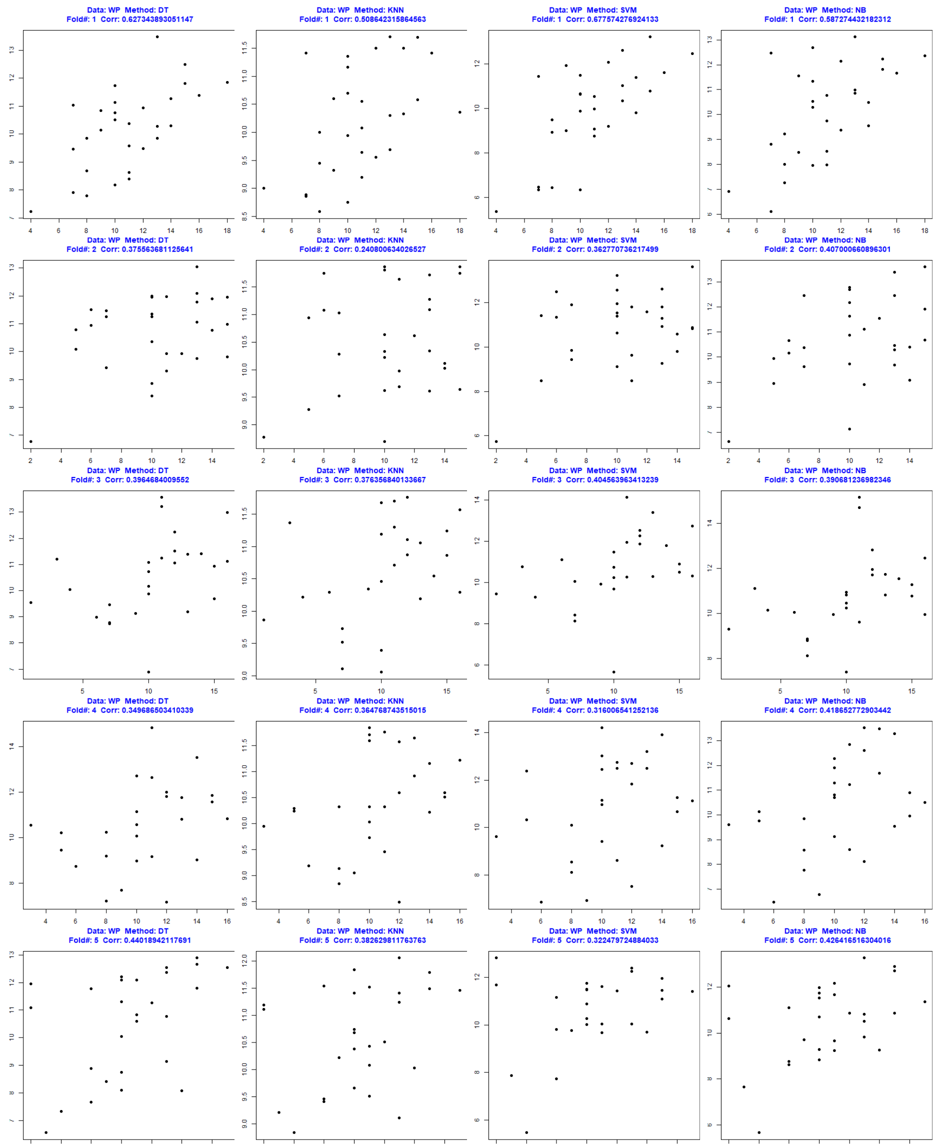


Fig. 9: Scatterplots of the overall perception of mental workload (x-axis), as reported by subjects and the prediction of the induced models (y-axis) for the 5 models produced by the regression algorithms (Extra trees: col 1; KNN: col 2; SVR: col 3; NB: col 4) employing the features of the Workload Profile