The Realtime Assesment of Mental Workload by Means of Multiple Bio-Signals

Masterthesis Report

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

Bart-Jan Boverhof, 6000142

Thesis Supervisor

Prof.dr.ir. B.P. Veldkamp

Date

October 14, 2020

1 Introduction

The topic of mental workload (MWL) has received widespread attention across a variety of different fields, amongst others the field of ergonomics (Young, Brookhuis, Wickens, & Hancock, 2015), human factors (Pretorius & Cilliers, 2007) and neurosciences (Shuggi, Oh, Shewokis, & Gentili, 2017). A simple definition of MWL is the demand placed upon humans whilst carrying out a certain task. As pointed out by de Waard (1996), such a definition is too shallow, for it defines workload solely in external sense. It is of importance to acknowledge that MWL is person-specific, for the amount of experienced MWL ushered by a given task may differ across people (De Waard & te Groningen, 1996). When referring to MWL throughout this research, person-specific MWL is meant specifically.

A commonly employed measure of workload is the NASA-Task Load Index (or TLX) questionnaire, operationalizing workload in clusters of six different dimensions (Hart, 2006). Such measurements are usually conducted post experiment, which could introduce some bias. An example of such a bias is the observer bias, prescribing that actors participating in an experiment tend to over-exaggerate the treatment effect when having to report it themselves post-experiment (Mahtani, Spencer, Brassey, & Heneghan, 2018).

An alternative method for assessing MWL is by measuring bio-signals during the experiment, and use these to classify the degree of perceived MWL. Examples of such bio-signals (hereafter modalities) include techniques such as electroencephalogram, eye tracking, galvanic skin response, functional near-infrared spectroscopy, etc. When training such classification models, administration of a questionnaire such as the TLX is usually utilized for labeling. The advantage of an approach like this is that complementary information streams, each stemming from a different modality can be interpreted simultaneously (Ramachandram & Taylor, 2017). This yields an objective and rich multifaceted classification of a mental construct, such as MWL. Additionally, a separate model for each respondent individually can be fitted, catering towards the individual perception of MWL. This approach comes however at the cost of an increase in complexity. This complexity resides in the construction of a complex framework that inputs the data from the utilized modalities, and outputs a single classification outcome.

The current research directly builds upon previous research conducted by Dolmans and colleagues, who proposed a framework for multi-modal deep-

learning classification of MWL (Dolmans, Poel, van 't Klooster, & Veldkamp, in press). The current research will explore the feasibility of a similar framework, but by utilizing different modalities and data. Additionally, the current research will aim to incorporate a real-time component. This implies that classification is done whilst the experiment takes place, enabling the possibilities to alter the state of the experiment in real-time. Such an approach could enable a wide range of possibilities: consider for example a flight-simulation with the objective of training pilots. The possibility for such a simulation to play out dynamically could enhance the learning-experience, namely by catering the state of the simulation towards the the individual, based on a real-time classification of for example their MWL.

A real-time classification approach will be contrasted with a non-real-time approach. Special focus will be placed upon the justification of using a real-time approach, given the increment in complexity such an approach is likely to cause. Ultimately, this line of research pursues the ability to conduct a dynamic experiment for multiple people simultaneously, and which can be altered by the researches in real-time.

In order to build both frameworks, three of the (arguably) most widely utilized modalities will be included. These modalities include the techniques of electroencephalogram (EEG), galvanic skin response (GSR) and photoplethysmography (PPG). It is important to stress that the objective of the current research is not to gain insight into the most optimal model for analyzing data stemming from the previously delineated modalities individually. It is rather to build a (real-time) framework researchers can utilize, and to which they can flexibly add modalities conform their own research goals. Consequently, one of two design principles on which the architecture of the framework reclines is the concept of modularity. Modularity refers the extent to which different modalities can freely be added and/or removed, without the necessity to re-architect and rebuild the entire framework. The second adhered design principle is the principle generalizability, prescribing that the framework should not solely be utilizable in the context of MWL, but also for the measurement of other mental constructs.

2 Methods

2.1 Related Work

The current section will provide an overview of previous research on the model design for each seperate modality. Additionally, the most feasible architecture for a multi-modular framework will be explored. In particular, attention is placed upon the data fusion, the real-time component and several model optimization techniques.

2.1.1 First modality: Electroencephalogram (EEG)

The first utilized modality is a technique called EEG, which detects electrical activity in the brain using electrodes. EEG is a widely utilized method for classifying MWL during experiments. With a review of the complete literature on EEG classification with deep learning, Craik and colleagues found a total of 16 % of all available papers to deal with MWL, lending credence to the widely employed phenomenon of EEG for investigating MWL (Craik, He, & Contreras-Vidal, 2019). Additionally, this review aimed to map the feasibility of using deep neural learning for a range of different EEG application separately. Summarizing the findings, Craik and colleagues report that studies mostly found deep belief networks and ConvNets to perform best when aiming to classify MWL, and advice one of these approaches consequently (Craik et al., 2019).

Research by Schirrmeister et al. (2017) contrasted several differently designed convolutional neural networks (ConvNets) against the baseline method for EEG data (FBCSP) in order to decode imagined or executed tasks. The investigated networks included deep, shallow, deep-shallow hybrid and residual ConvNets. Both the deep and shallow ConvNets were found to reach at least similar, and in some regard better classification results, as compared with the FBCSP baseline model. Altogether, a deep ConvNet with four convolutional-max-pooling blocks was found to perform best, displaying an accuracy of 92.4 % (Schirrmeister et al., 2017). The appropriate design choices were found to be of importance in order to be able to reach this accuracy, including the employment of techniques such as batch normalization, dropout and the usage of the ELU activation function.

A different approach is proposed by Tabar and Halici (2016), combining a ConvNet with a Stacked auto-encoder network (SAE). Within this network, the input layer feeds into a convolutional layer with the objective of learning the

filters and network parameters. The output of this convolutional layer subsequently feeds into SAE part of the network, constituting an input layer, 6 hidden layers and an output layer. A classification accuracy of 90 % was acquired with this model (Tabar & Halici, 2016) .

2.1.2 Second modality: Galvanic Skin Response (GSR)

The second utilized modality is GSR, measuring sweat gland on the hands and hereby inferring arousal. GSR readings have been found to significantly increase as a consequence of an increase in task cognitive load, hence constituting to be an objective predictor (Shi, Ruiz, Taib, Choi, & Chen, 2007).

Sun and colleagues explored the most optimal deep-learning model for classifying six different emotional states by means of GSR data. Several models were investigated, amongst others the support vector machine, the ConvNet, the long-short-term-memory (LSTM) model and a hybrid model combining the ConvNet and LSTM approaches. The hybrid model was found to perform best, exhibiting an accuracy of 74%. Additionally, data augmentation was found to be able to substantially increase classification results (Sun, Hong, Li, & Ren, 2019).

A variant on the CovNet LSTM model was employed by Dolmans et al. (in press), who aimed to classify MWL by means of amongst other modalities GSR (equally so as the current research). The performance of this model was contrasted with a network consisting solely of fully connected dense layers. Conform with findings by Sun and colleagues, the hybrid model was found to perform best (Dolmans et al., in press), displaying an accuracy of 82 %. The model architecture as utilized by Dolmans and colleagues deployed 2 convolutional layers, after each of which batch normalization was performed. Subsequently, max pooling was performed, after which the network feeds into two LSTM layers, followed by a dense layer.

2.1.3 Third modality: Photoplethysmography (PPG)

The third modality constitutes PPG, which is a technique utilized to measure heart rate. PPG is, not undeservedly, a widely deployed technique within the field of MWL classification. Zhang and colleagues investigated several approaches for measuring MWL, amongst three others the technique of PPG. Out of these four approaches, PPG was found to display both the highest sensitivity and reliability for measuring MWL, lending credence to the feasibility of PPG

as a method for classifying MWL (Zhang et al., 2018).

Work by Biswas and colleagues investigated a deep learning approach to PPG data, with the objective to perform both bio-metric identification and heart rate information. Exceptional results were attained with a neural network, attaining an average accuracy of 96 % (Biswas et al., 2019). This performance was managed with a hybrid model, incorporating two convolutional, followed with two LSTM layers. After each of two convolutional layers, batch normalisation and max pooling was applied.

The previously delineated model as proposed by Biswas and colleagues was adopted by Dolmans and colleagues, and subsequently applied towards the MWL case. This neural network was contrasted with a network of two fully connected dense layers. Batch normalisation and max pooling was performed in both contrasted networks. Surprisingly, the more sophisticated network constituting of the convolutional and LSTM layers was found to perform worse in the context of MLW classification, as compared with the simpler model, solely build from two dense layers (Dolmans et al., in press).

2.1.4 Fusion strategy

When architecting a multi-model framework, it is of importance to realize that the information stemming from the different modalities are required to be combined (i.e. fused) in order to ultimately result in a single classification. Fusion can be done at different time-points within the framework. Several fusion strategies as proposed by Ramachandram and colleagues will be considered (Ramachandram & Taylor, 2017).

Early (or data-level) fusion focuses on how to optimally combine data sources, before being fed into the classification model. Techniques that realize this include for example principle component analysis or factor analysis. Early fusing is usually challenging, especially in a multimodal scenario such as the current. This resides in the fact that data stemming from different modalities differ substantially in terms of dimensionality and sampling rate. Another disadvantage of early fusing, is that usually the oversimplified assumption of conditional independence is made. This assumption is unrealistic, for data stemming from different modalities are expected to be correlated in practice (Ramachandram & Taylor, 2017).

Late (or decision level) fusion on the other hand, refers to the process of aggregating the decisions from multiple models, each individually trained on all

modalities. In case the data sources stemming from the various modalities are either correlated or measured within a deviant dimensionality, late fusion is a much more feasible approach (Ramachandram & Taylor, 2017).

Lastly, intermediate fusion is the most widely employed fusion strategy for multi-modal deep-learning problems. Modalities are fused by simply adding a higher order layer, to which the individual deep-learning models separately defined for each modality feed into. This need not be a single layer, but could also be multiple layers, as long as each modality ultimately feeds into the highest order output layer. The depth of the fusion (i.e. the amount of fusion layers) can be chosen conform the specific situation, posing intermediate fusion to be the most flexible, and therefore the most widely used fusion method (Ramachandram & Taylor, 2017).

2.1.5 Real-time component

bla bla

2.1.6 Model optimization

The technique of batch normalization was proposed by Ioffe and Szegedy (2015), and is often applied in deep learning with the objective of enhancing the stability of a network. This is endeavored by including a batch normalization layer after each convolutional layer, re-centering and re-scaling the input feeding into this layer. If incorporating a batch normalization layer, it is recommended to do so before feeding into the activation function (Ioffe & Szegedy, 2015). An increase in accuracy for EEG classification was attained by Dolmans et al. (in press), Schirrmeister et al. (2017) by specifying a batch normalization layer after each convolutional layer. Equally so, the best performing model for PPG data as proposed by Biswas et al. (2019) included a batch normalization layer after each convolutional layer.

Pooling layers are often used in ConvNets, following a convolutional layer with the attempt to decrease the dimensionality. The objective of such layers are to merge similar features into one (for a more extensive elaboration see: (LeCun, Bengio, & Hinton, 2015)). Considering the EEG ConvNets, both Schirrmeister et al. (2017) and Tabar and Halici (2016) specified a max-pooling layer after each convolutional layer. The model proposed for GSR data by Sun et al. (2019) incorporate a max-pooling layer after each, but one of the convolutional layers. Lastly, the model as proposed for PPG data by (Biswas et al., 2019) specified

a max pooling layer after each convolutional layer.

Hyper-parameter optimization (HPO) is a technique that can be used to optimize hyper-parameter including learning rate, dropout probability and momentum. The Optuna toolbox provides a method for creating a parameter search space, from which values for the hyper-parameters can be sampled (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

2.2 Data

The current section will provide an overview of the utilized data. Special attention is placed on the experimental setup, the description of the participants and the data collection / synchronization process.

2.2.1 Experimental setup

The experimental setting for data collection is the spaceship videogame Empty Epsilon, in which the respondent is required to carry out tasks on a virtual spaceship (Daid & Nallath, 2016). This experiment is instituted by the Brain Computer Interfaces (BCI) testbed lab, hosted by the University of Twente (UT) and carried out in cooperation with Thales group Hengelo. The experiment constituted three different segments, during each of which respondents had to carry out different tasks, all aiming to measure workload. Each segment consists of six small sessions of roughly 5-10 minutes. These sessions varied in difficulty, including two easy, two intermediate and two hard sessions per segment. A schematic overview of the experimental structure is depicted as table 1. After each of the 18 segments, respondents filled in the TLX questionnaire consisting of 6 questions each, resulting in 18 filled in questionnaires. Each questionnaire inquired upon the degree to which the respondent experienced workload during the previous session. These ratings will be used as labels in later training. Within each segment, the order in which the sessions (varying in difficulty) were presented have been randomized. The order in which the segments were administrated were not random. Between every three sessions, respondents were requested to take a short 2 minute break.

The first segment emulated a scenario in which hostile spaceships approach the respondents spaceship. The respondent is required to quickly react, and defuse hostile spaceships in order to survive. The increment in difficulty caused the process of defusing hostile spaceships to be more challenging and hereby to take longer, aiming to increase workload consequently. The second segment em-

Table 1: Experimental setup: Number of sessions per segment and difficulty setting

	Easy	Intermediate	Hard
Segment 1	2	2	2
Segment 2	2	2	2
Segment 3	2	2	2

ulated a scenario in which the respondent had to navigate their spaceship trough space, gathering as many way-points as possible. Obstacles around which the respondent had to carefully navigate were introduced in the intermediate difficult scenario, and hostile spaceships the respondent had to decimate were introduced in the hard scenario. Both increase difficulty, and hereby aim to increase workload consequently. The third and final segment emulated a machine room, in which respondents had to control the power based on random requests generated by the video-game. Variables that could overheat the spaceship were introduced as a consequence of an increase in difficulty, demanding the respondent multi-task, hereby aiming to increase workload.

2.2.2 Participants

In total, twenty-five respondents are participating in the study. Currently, the data is still in the process of being collected, for which no additional descriptive statistics can be presented in the prevailing section. The respondents are students, recruited from the University of Twente. Recruitment has been conducted with Sona, which is a cloud-based participant management system. Requirements were that respondents didn't have constraints that might interfere with the utilized sensors, such as for example a pacemaker. This is assessed by means of a short demographic questionnaire prior to the experiment. Additionally, the respondents are made aware of ethical consent prior to the experiment, with the objective to ensure completely voluntary participation. Respondents were able to draw back from the experiment at any time.

2.2.3 Devices and Synchronization

The Shimmer3 GSR+ sensor is used for both PPG and GSR measurements. The device is worn on the wrist, and communicates the signals wirelessly. An ear-clip is utilized for measuring the PPG, and converting this to estimate heart

rate. Skin conductivity (GSR) is monitored between two electrodes attached to two fingers (Shimmer-Research, n.d.). Both PPG and GSR are measured on a sample rate of 256 Hz. EEG measurement is conducted with the shimmer 2, equally so on a sampling rate of 256 Hz.

Data streams stemming from the different modalities are measured on different frequencies, and consequently need to be properly synchronized. This is accomplished by means of an application called Lab-Streaming Layer (LSL), to which different data streams are streamed during the experiment. LSL properly synchronizes these data streams such that they refer to the same points in time, and records all data into a single file per participant (Kothe, Medine, & Grivich, 2018).

2.3 Framework Architecture

The architecture for the proposed framework is constructed by combining insights from the literature whilst keeping in mind the previously delineated design principles (i.e. the principles of modularity and generalizability). Figure # visualizes the entire framework *Still has to be made but didnt figure out yet how.*

The utilized network for the EEG modality is a ConvNet as proposed by Schirrmeister et al. (2017). The network is designed to include four convolutional layers, each follow by a max pooling layer of stride 3. The Exponential Linear Unit (ELU) function is utilized as activation function, defined as:

$$f(x) = \begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases} \tag{1}$$

The utilized network for the GSR modality is a LSTM ConvNet hybrid model, inspired by the work of Sun et al. (2019) and Dolmans et al. (in press). The network is designed to include two convolutional blocks, each consisting of a convolutional layer, followed by a batch normalization layer and closed with a max-pooling layer of stride 4. Following these blocks are two LTSM layers. The Rectified Linear Unit (ReLU) function is utilized as activation function, defined as:

$$f(x) = max(0, x) \tag{2}$$

Lastly, the PPG modality is analyzed by means of the model as proposed by Biswas et al. (2019). The nework starts with two convolutional blocks, each consisting of a convolutional layer, batch normalization layer and closed with a max pooling layer of stride 4. Following these blocks are two LTSM layers, equal to the GSR model. The utilized activation function is the ReLU, depicted as equation 2.

The network for each of the modalities are finally closed with one fully connected dense layer before feeding into the head network, such that all inputs are flattened in a lower dimensional space. All outputs from these dense layers are concatenated. The head network consists of Four dense layers, alternated with some batch normalization and max-pooling layers with the objective of stabilization. The entire network and its size, is summarized in table 2.

Table 2: Network summary

EEG	GSR	PPG	Head
Conv1: 25	Conv1: 128	Conv1: 128	Dense: 712
Conv2: 50	Conv2: 128	Conv2: 128	Dense: 356
Conv3: 100	LSTM1: 256	LSTM1: 256	Dense: 128
Conv4: 200	LSTM1: 256	LSTM2: 256	
Dense: 200	Dense: 256	Dense: 256	

Note that for convolutional layers the number reflects the amount of filters, whereas for all other layers it refers to the amount of nodes.

2.4 Model Evaluation

The performance of the real-time framework will be assessed by contrasting it with a regular, non-real-time framework. Specifically, the quality of performance for each modality individually will be compared across both frameworks. This validation will be endeavored by means of six widely used performance measures, all constructed from the confusion matrix, depicted as table 3

The measures accuracy, sensitivity, specificity, PPV, NPV and F1 will be utilized in order to asses model performance. The framework that performs best across these measures is considered to be the superior performing frame-

Table 3: Confusion matrix

	True Positive	True Negative
Predicted Positive	a	b
Predicted Negative	c	d

work. Table 4 depicts the constitution of these performance measures by partly referring to confusion matrix depicted as table 3.

Table 4: Performance Metrics

Accuracy:	$\frac{a+d}{a+b+c+d}$
Sensitivity:	$\frac{a}{a+c}$
Specificity:	$\frac{d}{b+d}$
Positive Predicted Value (PPV):	$\frac{a}{a+b}$
Negative Predicted Value:	$\frac{d}{c+d}$
F1-measure:	$\frac{2*Sensitivity*PPV}{Sensitivity+PPV}$

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2623–2631).

Biswas, D., Everson, L., Liu, M., Panwar, M., Verhoef, B.-E., Patki, S., ... others (2019). Cornet: Deep learning framework for ppg-based heart rate estimation and biometric identification in ambulant environment. *IEEE transactions on biomedical circuits and systems*, *13*(2), 282–291.

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, *16*(3), 031001.

Daid, & Nallath. (2016). *Empty epsilon multiplayer spaceship bridge simulation.* URL: https://github.com/daid/EmptyEpsilon. GitHub.

De Waard, D., & te Groningen, R. (1996). *The measurement of drivers' mental workload*. Groningen University, Traffic Research Center Netherlands.

Dolmans, T., Poel, M., van 't Klooster, J.-W., & Veldkamp, B. (in press). Percieved mental workload detection using intermediate fusion multi-modal networks.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kothe, C., Medine, D., & Grivich, M. (2018). Lab streaming layer (2014). *URL: https://github. com/sccn/labstreaminglayer (visited on 02/01/2019)*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Mahtani, K., Spencer, E. A., Brassey, J., & Heneghan, C. (2018). Catalogue of bias: observer bias. *BMJ Evidence-Based Medicine*, *23*(1), 23.

Pretorius, A., & Cilliers, P. (2007). Development of a mental workload index: A systems approach. *Ergonomics*, *50*(9), 1503–1515.

Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., ... Ball, T. (2017). Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, *38*(11), 5391–5420.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In *Chi'07 extended abstracts on human factors in computing systems* (pp. 2651–2656).

Shimmer-Research. (n.d.). *Shimmer3 gsr unit.* Retrieved from `https://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor`

Shuggi, I. M., Oh, H., Shewokis, P. A., & Gentili, R. J. (2017). Mental workload and motor performance dynamics during practice of reaching movements under various levels of task difficulty. *Neuroscience*, *360*, 166–179.

Sun, X., Hong, T., Li, C., & Ren, F. (2019). Hybrid spatiotemporal models for sentiment classification via galvanic skin response. *Neurocomputing*, *358*, 385–400.

Tabar, Y. R., & Halici, U. (2016). A novel deep learning approach for classification of eeg motor imagery signals. *Journal of neural engineering*, *14*(1), 016003.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.

Zhang, X., Lyu, Y., Hu, X., Hu, Z., Shi, Y., & Yin, H. (2018). Evaluating photoplethysmogram as a real-time cognitive load assessment during game playing. *International Journal of Human–Computer Interaction*, *34*(8), 695–706.