Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

MSc Thesis Bart-Jan Boverhof (6000142)
Physiological sensor based prediction of mental workload: A multimodal deep learning approach
May 2021

Supervisor:
Prof. Dr. Ir. Bernard Veldkamp

Second grader:
Prof. Dr. Ir René Eijkemans

Preferred journal of publication: Frontiers in Neuroscience

Physiological sensor based prediction of mental workload: A multimodal deep learning approach

Bart-Jan Boverhof *[1]

[1]Faculty of Social and Behavioural Sciences, Utrecht University

ABSTRACT

The assessment of mental workload by means of physiological sensors has gained in popularity in recent years. We utilized three physiological sensors from which mental workload was predicted, being electroencephalography, photoplethysmography and galvanic skin response. Deep learning poses a promising, yet novel, approach towards the modeling of such sensory data. Four deep neural network architectures were created and contrasted in their performance. Three out of four were specified to predict workload based on on a single of the aforementioned physiological sensors, whilst the fourth architecture was a joint neural network combining all physiological signals into a single model. In this study we aimed to identify 1) which physiological sensor constitutes the most adequate predictor of mental workload, and 2) whether a joint approach utilizing multiple sensors is preferable over the approaches utilizing just one. Altogether the electroencephalography-only architecture was found to perform best, closely followed by the galvanic skin response-only architecture, both displaying a mean absolute error of 0.110 and 0.119 (scaled on 0-1) respectively. Against expectations, the joint approach was outperformed by both aforementioned architectures, displaying a mean absolute error of 0.128 (equally so scaled on 0-1). The photoplethysmography-only architecture was found to perform worst, displaying a mean absolute error of 0.136. In the discussion section we presented several explanations that help rationalize this anomaly, one being due to an imperfect model architecture for the joint neural network, and another being due to overfitting resulting from a limited amount of data. Regardless of the unexpected character of this result, we still recognize that differences in performance were of rather modest size: all four neural network architectures performed fairly adequately.

1 Introduction

The topic of mental workload is a widely studied phenomenon across a variety of different fields, amongst others the field of ergonomics (Young, Brookhuis, Wickens, & Hancock, 2015), human factors (Pretorius & Cilliers, 2007) and cognitive neurosciences (Shuggi, Oh, Shewokis, & Gentili, 2017). A commonly utilized definition of mental workload, hereafter referred to as simply "workload", is the demand placed upon one whilst carrying out a particular task. As rightfully pointed out by de Waard and te Groningen (1996), the aforementioned definition is incomplete, for it defines workload solely as a phenomenon external to the individual. Workload, however, requires to be recognized as a person-specific construct, for the amount of perceived workload ushered by a given task may vary substantially between individuals (de Waard & te Groningen, 1996).

A commonly employed method for the assessment of workload in an experimental setting is the well established NASA-Task Load Index questionnaire, hereafter referred to as “NASA-TLX”. This method embodies several dimensions, jointly inquiring into the amount of perceived workload (Hart, 2006). However, due to the ever-increasing availability in computing power characterizing recent times, a novel approach to workload assessment has gained in popularity. This promising approach constitutes the use of physiological sensors, worn throughout the entire duration of the experiment. An approach like this poses the advantage that the development of perceived workload may be monitored throughout the entire experiment, enabling the possibility to precisely monitor the impact of a specific task on a participant’s level of workload. Another advantage is that the method may be personalized, which could simply be done by creating and training an individual model for each participant. By doing this, the

*b.boverhof@students.uu.nl

person-specific character that is inherent to workload is acknowledged, the importance of which was stressed in the opening paragraph of this paper.

With the current research we explore the feasibility of an approach towards workload assessment by means of various physiological sensors, hereafter referred to as “modalities”. The three utilized modalities are electroencephalography (EEG), photoplethysmography (PPG) and galvanic skin response (GSR). EEG is a technique that monitors electrical activity in the brain, and is often utilized in the assessment of workload (Berka et al., 2005; Craik, He, & Contreras-Vidal, 2019). EEG is found to be amongst the most adequately performing techniques for this task (Hogervorst, Brouwer, & Van Erp, 2014). GSR is a technique that measures electrical conductance of the skin, and equally so is a widely adopted method for workload assessment (Nourbakhsh, Wang, Chen, & Calvo, 2012; Zhou, Jung, & Chen, 2015). Lastly, also heart-rate is a widely utilized indicator of workload, mostly obtained through PPG (Zhang et al., 2018; Jimenez-Molina, Retamal, & Lira, 2018).

In this project we opted for a deep learning approach towards workload prediction. A range of four different deep neural network architectures, hereafter referred to as “DNN’s” or simply “networks”, are constructed and contrasted in their performance. Among these are three unimodal DNN’s, i.e. networks each utilizing only a single modality. Thus, one network solely utilizes the EEG modality, one solely utilizes the GSR modality and one solely utilizes the PPG modality. The fourth is a multimodal network, i.e. combining all three aforementioned modalities into a single architecture. The main advantage of taking a deep learning approach specifically resides in its flexibility. Fitting a model that combines multiple modalities can be done in different, relatively straightforward, ways when taking a deep learning approach (Ramachandram & Taylor, 2017). We expect the multimodal approach to results in better performance as compared with the unimodal approaches, for predictions are based on richer composite of information. This expectation is indeed in line with previous research, which reported a multimodal approach to perform superiorly (Dolmans, Poel, van ’t Klooster, & Veldkamp, 2020; Han, Kwak, Oh, & Lee, 2020; Rastgoo, Nakisa, Maire, Rakotonirainy, & Chandran, 2019; Yin, Zhao, Wang, Yang, & Zhang, 2017).

With this research we 1) seek to explore which individual modality constitutes the most adequate predictor of workload when taking a deep learning approach, and 2) whether a multimodal network architecture is preferable over the three simpler (but much less computationally demanding) unimodal architectures. This comparison will mainly be propelled by contrasting network performance, however when drawing conclusions, considerations regarding computational costs will also be taken into consideration. With this research we aim to deepen the knowledge of taking a deep learning approach towards workload prediction with physiological sensors in an experimental setting, which as of yet, is still a novel and fairly understudied topic.

2 Data & Methods

2.1 Related Work

The following section provides an overview of preceding research on the assessment of workload by means of various physiological sensors, with a focus on deep learning approaches. Attention is predominantly placed upon network architectures and hyperparameter optimization. In this section we will solely describe previous work: an in depth description of our adopted modeling approach follows in Section 2.3.

EEG Network Architecture

An overview of the complete literature EEG applications with deep learning was presented by Craik et al. (2019), who reported 16% of all publications to constitute workload assessment. The lion’s share of these publications utilized either a deep belief network or convolutional neural network (ConvNet). Tabar and Halici (2016) proposed a ConvNet and stacked auto-encoder (SAE) hybrid approach towards workload classification with EEG. Their proposed network constituted two different parts. Firstly, raw EEG data was fed into a single convolutional layer, with the objective of extracting features. Secondly, the the extracted features fed into the SAE part of the network, which encompassed a stack of dense-layers designed to ultimately result in a classification of workload. An accuracy of 90% was acquired with this architecture (Tabar & Halici, 2016). A similar approach was proposed by Schirrmeister et al. (2017), who contrasted different DNN architectures in their ability to adequately decode EEG signals. Several ConvNet architectures were contrasted against the baseline method for EEG decoding, filter bank common spatial pattern (FBCSP). A deep ConvNet, a shallow ConvNet, a deep-shallow hybrid ConvNet and a residual ConvNet were contrasted with the FBCSP baseline method. Both the deep and shallow ConvNets were found to reach at least similar, and in some regards better classification results as compared with the FBCSP baseline approach. Altogether, a deep ConvNet with four convolutional-max-pooling blocks was found to perform best, exhibiting a classification accuracy of 92.4% (Schirrmeister et al., 2017).

GSR Network Architecture

Sun, Hong, Li, and Ren (2019) explored various deep-learning approaches towards the classification of emotional states with GSR. Despite that this application differs from workload prediction, inspiration may still be drawn from their utilized architecture. Various models were explored, amongst others a support vector machine, a ConvNet and a long-short-term-memory (LSTM) network. In addition to these, a hybrid ConvNet-LSTM approach was explored as well. This hybrid approach constituted four convolutional blocks followed by a single LSTM layer, and was found to perform best with an accuracy of 74% (Sun et al., 2019). Dolmans et al. (2020) contrasted several physiological signals in their ability to predict workload, each of which by taking a deep learning approach. One of these modalities constituted GSR, for which a ConvNet-LSTM hybrid network architecture was utilized to predict workload. The performance of this DNN was contrasted with a network consisting solely of fully connected dense layers. Conform with the previously mentioned findings by Sun et al. (2019), the ConvNet-LSTM hybrid architecture was found to perform best, displaying an absolute difference between predicted and true label of 0.180 (scaled on 0-1). This network architecture incorporated two convolutional blocks, followed by two LSTM layers (Dolmans et al., 2020).

PPG Network Architecture

Research by Biswas et al. (2019) explored a PPG deep learning approach, in which their objective was to perform both bio-metric identification and obtain heart rate information. Again, despite that this research doesn't constitute workload prediction or classification, inspiration may still be drawn from the employed network architecture. An exceptional performance of 96% accuracy was realized with a ConvNet-LSTM hybrid, incorporating two convolutional max-pooling blocks followed by two LSTM layers (Biswas et al., 2019). Dolmans et al. (2020) also explored a deep-learning PPG approach to workload prediction. A ConvNet constituting two convolutional blocks was found to perform best out of all contrasted PPG architectures, displaying an absolute difference between predicted and true label of 0.1969 (scaled on 0-1).

Multimodal Fusion Strategy

When creating a DNN that combines multiple modalities into a single network, the information streams that stem from the different subparts of the network (each corresponding to one of the modalities) are required to be combined, i.e. "fused", at a given point in the network. This is necessary in order to ultimately end up with a single prediction of workload. Fusion can be done conforming several strategies, out of which three strategies as proposed by Ramachandram and Taylor (2017) are considered.

Firstly, early fusion constitutes an approach that fuses data sources before being fed into the network. Early fusing usually proves to be quite challenging, residing in the fact that data streams stemming from different modalities often differ in their dimensionality and sampling rate. In addition, when taking an early fusion approach, the assumption of conditional independence is made implicitly. This assumption is oftentimes unrealistic in practice, for data stemming from different modalities are expected to be correlated (Ramachandram & Taylor, 2017). Secondly, late fusion refers to the process of aggregating the decisions of multiple separate networks, each constructed for every modality separately. When the data sources stemming from the various modalities happen to be either correlated or different in their dimensionality, late fusion is oftentimes a more feasible approach as opposed to early fusion (Ramachandram & Taylor, 2017). Lastly, intermediate fusion is the most widely employed fusion strategy for multimodal networks. Data streams are usually fused by a concatenation layer, joining the outputs of the separately defined network parts of each modality. This results in a single joint deep neural network. Several higher-order layers are usually defined in between the concatenation layer and the ultimate prediction layer. The depth of the fusion, i.e. the specified number of higher-order layers, may be chosen by the researcher, posing intermediate fusion to be the most flexible and therefore the most widely adopted fusion strategy (Ramachandram & Taylor, 2017).

Indeed when consulting the literature, it becomes apparent that intermediate fusion is the most prevailing strategy for multimodal DNN architectures. As mentioned before, when taking an intermediate fusion approach, the higher-order part of the network is required to be designed by the researcher. Rastgoo et al. (2019) utilized a multimodal ConvNet approach, and fused the modalities by concatenation, followed with two LTSM layers and two dense layers. A simpler approach is adopted by Han et al. (2020), who utilized an intermediate fusion approach solely consisting of several fully connected dense layers. Lastly, Dolmans et al. (2020) took a relatively deep intermediate fusion approach, consisting of two dense layers, two convolutional layers followed by another two dense layers.

Auxiliary Layers in a DNN Architecture

At the core of a DNN are the aforementioned types of layers, such as convolutional, LSTM and dense layers. Such layers are often accompanied with auxiliary layers in order to ensure network stability. The first type of layers often used in network architectures are batch normalization layers, which have the objective improving network stability. It is beneficial to incorporate a batch normalization layer subsequent to a convolutional layer, but before feeding into the activation layer (Ioffe & Szegedy, 2015). For a deep dive into batch normalization, we refer the reader to Ioffe and Szegedy (2015). All previously considered network architectures made use of batch normalization layers (Biswas et al., 2019; Dolmans et al., 2020; Schirrmeister et al., 2017; Sun et al., 2019; Tabar & Halici, 2016).

Pooling layers are commonly employed in ConvNets, usually succeeding a convolutional layer with the purpose of reducing dimensionality. The objective of such layers are to down-sample features into a more compact space, hereby only retaining essential information, and thus omitting redundant information. For a more extensive elaboration on pooling we refer the reader to LeCun, Bengio, and Hinton (2015). All previously considered network architectures utilize max-pooling layers, usually specified subsequent to the activation layer (Biswas et al., 2019; Dolmans et al., 2020; Schirrmeister et al., 2017; Sun et al., 2019; Tabar & Halici, 2016).

The incorporation of dropout layers was initially proposed by Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014). Dropout layers drop a randomly-selected number of neurons and their connections throughout the training process. More specifically, within each full training run through the data, i.e. each “epoch”, a different randomly selected set of neurons are fixed to be inactive. Doing so helps to reduce overfitting, particularly for complex network architectures (Srivastava et al., 2014). Most of the previously considered network architectures make use of dropout layers (Biswas et al., 2019; Dolmans et al., 2020; Schirrmeister et al., 2017; Sun et al., 2019).

Hyperparameter Optimization

Hyperparameter optimization (HPO) is a technique that optimizes network hyperparameters, such as for example learning rate and dropout rate. Substantial advancements in DNN performance have been attained by utilizing HPO, especially in the case of ConvNets (Bergstra & Bengio, 2012). In theory, one could optimize the entire DNN architecture, including the number of neurons/filters in dense/convolutional layers, the number of layers in general, whether to use certain layers, etc. There is, however, a strong relationship between the amount of parameters to be optimized and the computational resources required to do so: optimizing many hyperparameters will inflate computational costs substantially. The Optuna toolbox provides a method for creating a parameter search space, from which values for the hyperparameters may be sampled, and optimization can be performed (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

2.2 Data

The subsequent section provides a description of the utilized data and its collection process. Attention is placed upon the experimental setup, the description of the respondents, the utilized devices for data collection, the synchronization process and the data preprocessing approach.

Experimental Setup & Data Labeling Strategy

The experimental setting in which our data was collected is the open-source spaceship video-game Empty Epsilon, in which partakers carried out tasks on a virtual spaceship bridge (Daid & Nallath, 2016). This experiment was instituted by the Brain Computer Interfaces Testbed [1] hosted by the University of Twente and carried out in cooperation with Thales Group[2] Hengelo, the Netherlands. The experiment was divided into a total of 15 tasks, differing in difficulty and hence designed to evoke varying degrees of workload. The total duration of the experiment was roughly three hours.

Respondents were asked to fill out the NASA-TLX questionnaire after each of the fifteen task, inquiring into the amount workload perceived during the respective task. Each questionnaire consisted out of six items. The mean of a subset of four of these items were utilized for labeling of the data. Two items were not incorporated into the scale, for they inquired into the perceived physical demands. Since the experiment took place on a computer whilst participants sat behind a desk, we chose to not incorporate such items. Each item, and hence the newly constructed scale, ranges from 0-20, wherein 0 reflects the lowest possible level workload and 20 reflects the highest possible level workload. In order to encourage numerical stability, label scores were normalized to reside in-between 0-1.

Participants

After having to omit seven participants due to hardware failure, 27 respondents remained in the final analysis. 18 participants were female whereas nine were male. The average age and its standard deviation were $\mu = 26$ and $\sigma = 10.31$ respectively. The participants were students recruited from the University of Twente situated in the Netherlands, as well as several employees of Thales Group Hengelo, equally so situated in the Netherlands.

Devices and Synchronization

The Shimmer3 GSR+ sensor was used for both PPG and GSR measurements. This device was worn on the wrist, whilst signals were communicated wirelessly via Bluetooth. An ear-clip was utilized for measuring PPG, which was automatically converted to heart-rate. Skin conductivity, or GSR, was measured by two electrodes attached to the fingers. EEG measurement was conducted with the Muse 2 multi-sensor headband, equally so whilst communicating signals wirelessly via Bluetooth. The Shimmer3 GSR+ measured on a sampling rate of 256 Hz, whereas the Muse 2 measured at a sampling rate of 220 Hz.

Given that multiple devices were utilized, data streams were required to be properly synchronized. This was accomplished by means of an application called Lab-Streaming Layer, hereafter referred to as "LSL", developed by Kothe, Medine, and Grivich (2018). The three data streams stemming from the two devices were all streamed to LSL during the experiment. LSL subsequently properly synchronizes all data streams in real-time, such that they are parallel, hence all referring to equivalent points in time.

Data Preprocessing & Partitioning

For each participant, the EEG, PPG and GSR data streams were cut into windows of eight seconds, resulting in an average of roughly 351 windows per respondent. 80% of all windows have been used for training, 10% for validation within the training process and 10% for the testing of network performance. The partitioning of the windows into training, validation and testing sets was done systematically by selecting every *n-th* window, such that windows over the entire duration of the experiment were represented equally in each partition. Exactly the same partitions were utilized in the training, validation and testing procedure across the four architectures to ensure a fair comparison.

2.3 Network Architectures & Training

Individuals tend to respond differently towards high workload situations: for some workload manifests itself mostly through an increase in heart rate detected through PPG, whereas for others it manifests as an increase in skin conductivity detected through GSR. To account for such between-person differences, we trained an individual DNN for each participant. Since a total of four architectures are contrasted for 27 respondents, this resulted in a total of 108 trained networks. Equally so, HPO was conducted for each person individually, resulting in 108 sets of optimized

[1] https://bmslab.utwente.nl/

[2] https://www.thalesgroup.com/en/countries/europe/netherlands

hyperparameters. All networks were constructed in Python, by utilizing the deep-learning toolbox PyTorch (Paszke et al., 2017).

Unimodal Network Architectures

All unimodal architectures drew inspiration from aforementioned research (Biswas et al., 2019; Dolmans et al., 2020; Schirrmeister et al., 2017; Sun et al., 2019). We opted for a ConvNet architecture for all three unimodal network architectures.

The three unimodal DNN architectures are depicted alongside one another in Figure 1. Each of the unimodal networks may be distinguished by two separate parts, being the convolutional part and the prediction part. The objective of the convolutional part is to extract features from the data, and consists of four convolutional blocks, each constituting a convolutional layer, a batch normalization layer, the Exponential Linear Unit (ELU) activation function, and a max-pooling layer respectively. The objective of the prediction part is to predict the amount of perceived workload, i.e. ultimately resulting in a single prediction of the label. The prediction part constitutes a flatten-layer, with the objective of representing all input into one dimensional shape, followed by a dropout layer, a dense layer, a Rectified Linear (RELU) activation function closed by another dense layer. An overview of the amount of utilized filters / neurons for each layer of each network is provided in Table 1.

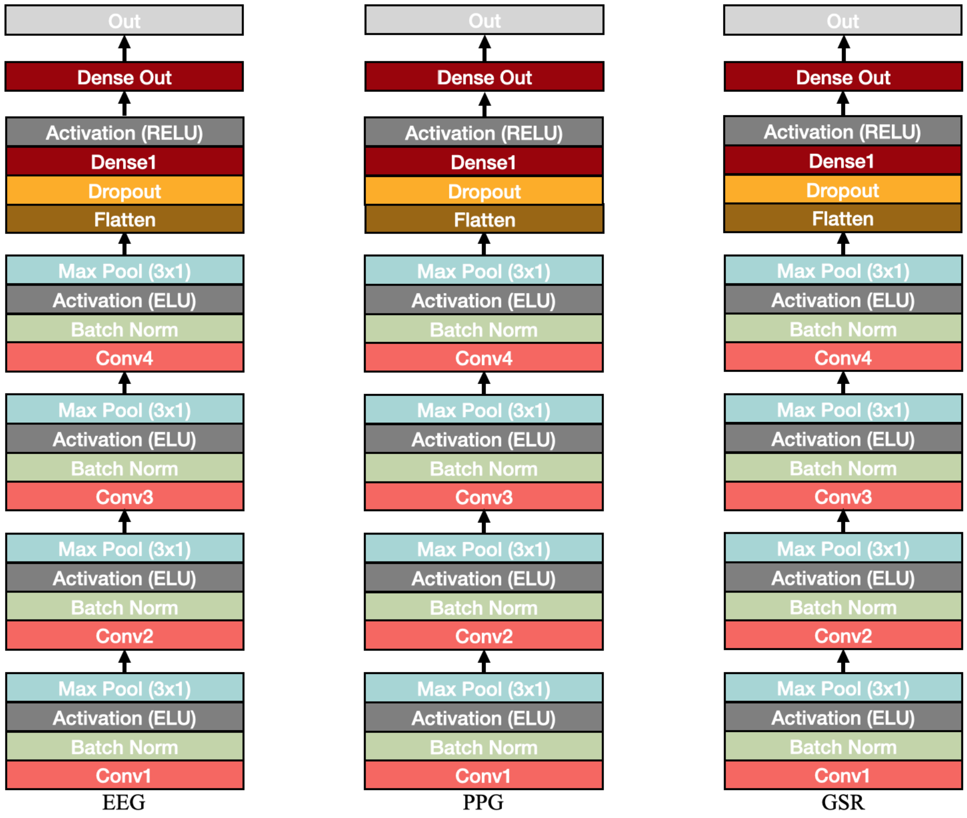

Figure 1. Unimodal Network Architectures. Each network constitutes a convolutional part of four convolutional blocks, and a classification part of two fully connected dense layers.

We optimized the hyperparamater for learning rate, dropout rate and the amount of neurons for the first dense layer ("Dense1" in Figure 1). This was managed in a total of 50 trials per network. With respect to network training, a total of 600 epochs have been made through the data for each network specifically. All HPO and training of the unimodal networks was done on a V100 GPU, offered by the GPU cloud service Google Collab. Total running time for the unimodal networks was about 96 hours, the most of which was absorbed by HPO.

Table 1. Amount of neurons and filters that each respective layer outputs.

	EEG-Net/Part	GSR-Net/Part	PPG-Net/Part	Multimodal Head-Net
Unimodal-Nets	Dense Out: 1	Dense Out: 1	Dense Out: 1	
	Dense 1: *hpo*	Dense1: *hpo*	Dense1: *hpo*	
	Dropout: *hpo*	Dropout: *hpo*	Dropout: *hpo*	
	Conv4: 200	Conv4: 128	Conv4: 128	
	Conv3: 100	Conv3: 64	Conv3: 64	
	Conv2 50	Conv2: 32	Conv2: 32	
	Conv1: 25	Conv1: 16	Conv1: 16	
Multimodal-Net	Dense EEG1: *input*	Dense PPG1: *input*	Dense GSR1: *input*	Dense Out: 1
	Conv EEG4: 200	Conv PPG4: 128	Conv GSR4: 128	Dense3: *hpo*
	Conv EEG3: 100	Conv PPG3: 64	Conv GSR3: 64	Dense2: *hpo*
	Conv EEG2 50	Conv PPG2: 32	Conv GSR2: 32	Dropout: *hpo*
	Conv EEG1: 25	Conv PPG1: 16	Conv GSR1: 16	

For all convolutional layers, the depicted numbers reflect the amount of filters that are outputted by the respective layer, whereas for dense layers it reflects the amount of nodes outputted by the respective layer."Input" refers layers wherein the number of outputting nodes equals the number of inputting nodes.

Multimodal Network Architectures

The multimodal network architecture was determined by combining the previously characterized unimodal architectures into a single architecture. These separate network parts were joined by a head-network, the architecture of which drew inspiration from aforementioned research by Han et al. (2020). A visual representation of the multimodal network is depicted in Figure 2. An overview of the amount of utilized filters and neurons is provided in Table 1.

The feature extraction part of the multimodal network architecture is highly similar to the separate unimodel architectures. The seperate unimodal parts were combined by means of an intermediate fusion strategy, due to its highly flexible nature as compared with alternative fusion strategies (See: Section 2.1). The outputs of the three distinct subparts were flattened, followed by a dense layer and RELU activation. Flattening was necessary such that all inputs were represented into one-dimensional space, after which concatenation rendered possible. The prediction part, also referred to as "Head-Network", consists of three dense layers, alternated with a dropout layer, batch normalization layers and RELU activation layers.

We optimized the hyperparameter for learning rate, dropout rate and the amount of neurons for the first and second dense layers in 18 trials per network. For training, a total of 250 epochs have been made through the data for each network independently. All training and HPO has been done on a Cloud TPU v2, offered by the GPU cloud service Google Collab Pro. Total running time for the multimodal networks was about 7 days, the most of which absorbed by HPO.

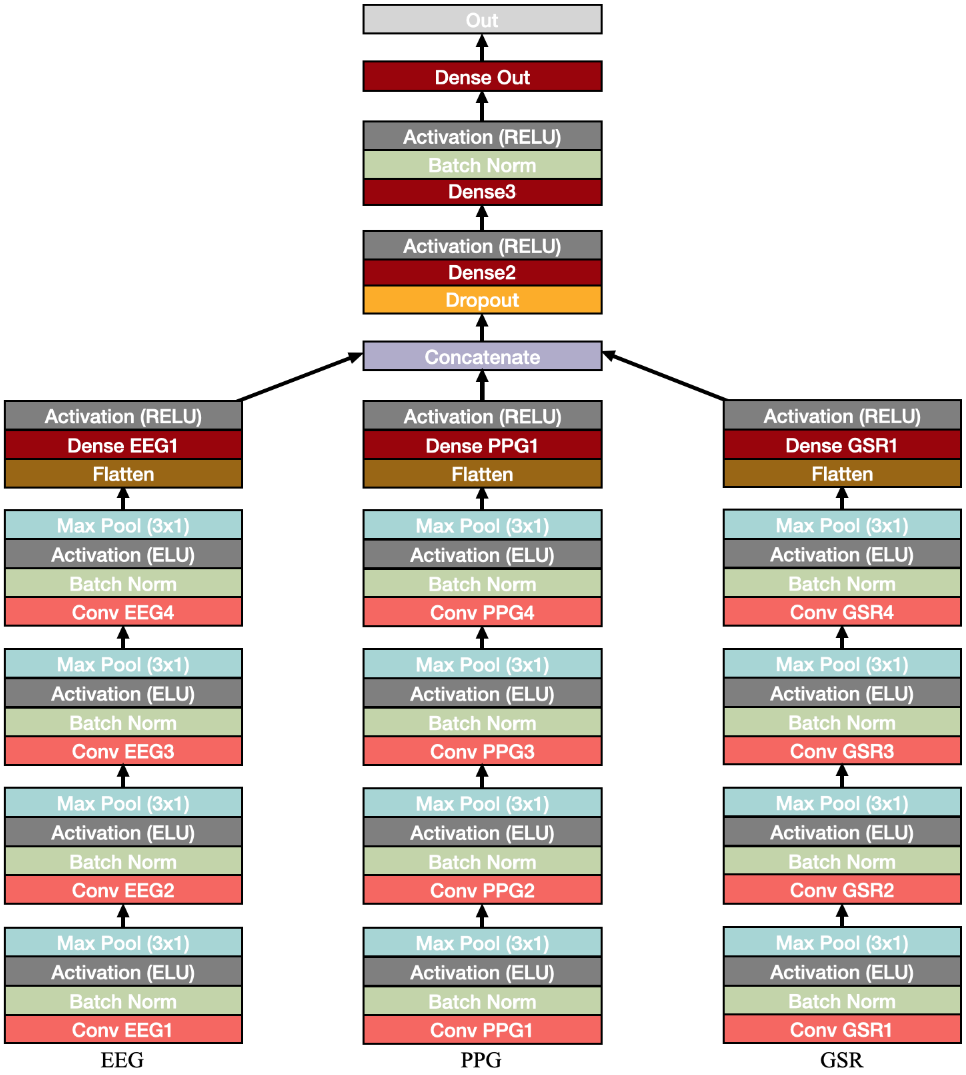

Figure 2. Multimodal Network Architecture. The network constitutes a convolutional part of four convolutional blocks for each of distinct network parts. All three distinct parts were flattened and concatenated subsequently. The prediction part of the network constitutes several dense layers.

2.4 Network Performance Evaluation

For each of the four investigated network architectures, the predictions of all 27 individually trained network were aggregated, after which performance was assessed by means of the following three performance metrics.

The Mean Absolute Error, hereafter referred to as "MAE", is the first utilized performance metric, defined as:

$$\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-y_i|$$

where n refers to the total amount of testing windows, $f(x_i)$ to the predicted value for window i and y_i to the true label of window i. The MAE constitutes the most straightforward metric for DNN performance assessment, for its value can simply be interpreted as the average misprediction.

The Root Mean Squared Error, hereafter referred to as "RMSE", is the second utilized performance metric, defined as:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(f(x_i)-y_i\Big)^2}$$

where n refers to the amount of testing windows, $f(x_i)$ to the predicted value for window i and y_i to the true label of window i. The RMSE is closely related to the MAE, differing in that it tends to punish big differences more severely.

Finally, Pearson's correlation coefficient, hereafter referred to as simply "correlation", constitutes the third utilized performance metric, defined as:

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where X refers to the predicted test windows, Y to the true labels of the test windows, $cov(X,Y)$ to their covariance and σ_X and σ_Y to the standard deviation of x and y respectively. Conceptually, correlation can be interpreted as the strength of the linear relationship between predictions and their labels, thus providing an indication of network performance.

3 Results

Before addressing network performance, a description of all window labels is depicted in Figure 3. This constitutes an aggregation of the labels of all windows over all participants, resulting in a total of 9474 depicted window labels. It becomes readily apparent that the distribution of the window labels displays a noticeably right-skewed tendency, with a median of 5.75 and an 90% upper-bound quantatile of 12.75. This implies that there is a considerable underrepresentation of high workload windows.

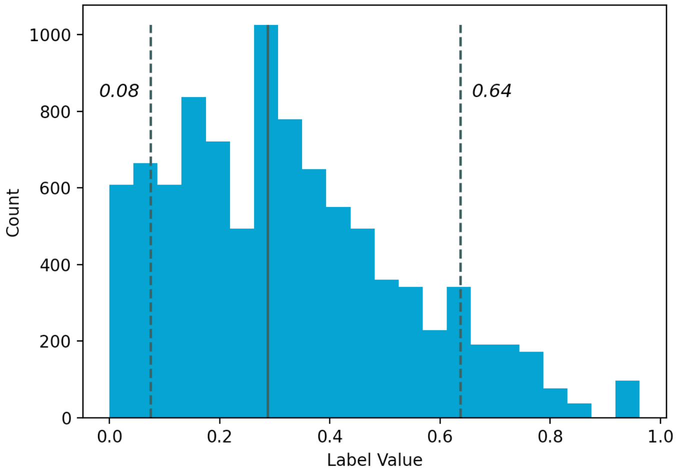

Figure 3. Distribution window labels. The solid line represents the median, whereas the two dashed line and their displayed values represent the 10% and 90% quantiles respectively.

3.1 Deep Neural Network Performance

Depicted in Table 2 are the performance metrics for each of the four network architectures. The most remarkable result is the relatively bad performance of the multimodal architecture. Both the unimodal EEG and GSR architectures score consistently higher on all metrics.

Performance in terms of MAE is most favorable for the EEG architecture, closely followed by the GSR architecture. The multimodal architecture performs worse as compared with the aforementioned architectures, ultimately followed by the PPG architecture which performed worst. Results in terms of RMSE adhere to a similar pattern, which is not surprising given that the RMSE and MAE are closely related. RMSE values are consistently higher as compared with MAE, which can be attributed to the tendency of the RMSE to punish strong deviations more severely. The correlation between predictions and labels are indicative of a similar pattern, in which the EEG architecture scores best, followed by the GSR, multimodal and PPG -architectures respectively Despite consistency in this pattern, an interesting disparity is that the differences in terms of correlation between the architectures are more substantial. In addition to the aforementioned results, we also assessed network performance on the training data, with the objective of gaining insight into overfitting. Table A.1 in Appendix A depicts the performance metrics of the networks as deployed on the training data. It becomes apparent that the unimodal architectures perform substantially better on the training data, whereas the multimodal architecture performs slightly better on the training data.

Table 2. Deep Neural Network Performance Metrics

	MAE (sd)	MAE (sd)	RMSE	Pearson's
	0-1 scale	0-20 scale	0-1 Scale	Correlation
EEG	0.110 (0.15)	2.206 (3.08)	0.154	0.688
PPG	0.136 0.18)	2.717 (3.59)	0.180	0.530
GSR	0.119 (0.16)	2.377 3.18)	0.159	0.653
Multimodal	0.128 (0.17)	2.570 (3.33)	0.167	0.609

Values between parentheses are standard deviations. MAE refers to Mean Average Error. RMSE refers to Root Mean Square Error. The depicted values are the averages of all individually trained networks.

An independent samples t-test was conducted to formally test differences in terms of absolute (window) error across various pairs of architectures. The absolute error of a window i is defined as:

$$|f(x_i) - y_i|$$

where $f(x_i)$ represents the predicted value for window i and y_i represents its true label. Testing was done two-sided with $\alpha = 0.05$. Mean absolute error values per architecture are not reported, for they can be discerned from Table 2. The absolute error for the unimodel EEG architecture was found to be significantly lower as compared with the multimodel architecture $t(1936) = -2.566, p = 0.01$. No significant difference in terms of absolute error between the EEG- and GSR -architectures was found $t(1936) = -1.188, p = 0.24$, and equally so no significant difference was found between the GSR- and multimodal -architectures $t(1936) = -1.463, p = 0.144$. Lastly, absolute error for the GSR architecture was found to be significantly lower as compared with the PPG architecture $t(1936) = -2.636, p = 0.008$, but no significant difference was found between the GSR- and multimodal -architecture $t(1936) = -1.164, p = 0.245$.

To explore performance differences across architectures into more detail, a graphical visualization of the distribution of window prediction errors per architecture are depicted in Figure 4.

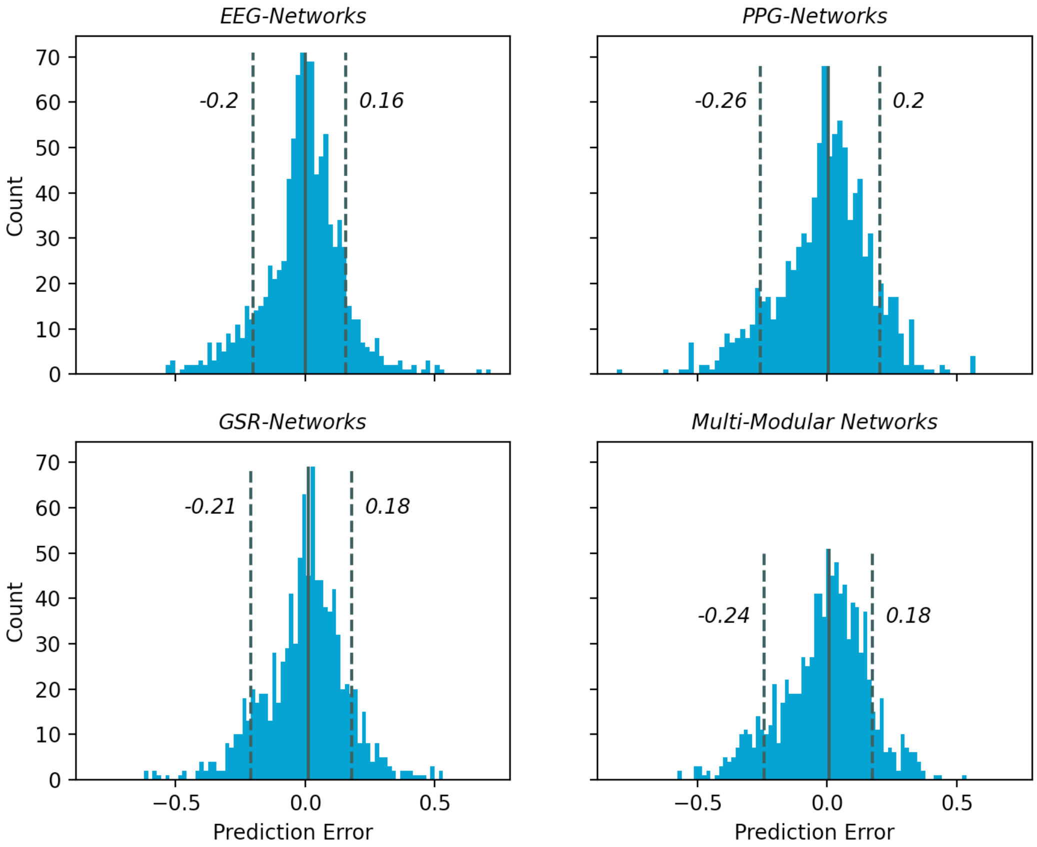

Figure 4. Distribution of window prediction error per network architecture. The solid lines represent the median, whereas the dashed lines represent the 10% and 90% quantiles respectively.

In line with the previously presented performance metrics, it becomes readily apparent that the distribution of the prediction errors for the architecture that scored highest across all metrics, i.e. the EEG architecture, approximates the normal distribution most closely. This may be derived by the strong accumulation around 0 error (i.e. visually "peaked") and the relatively flat tails. The two worst performing architectures, i.e. multimodal- and PPG -architectures, display relatively thick tails, and are observed to have less strong accumulation around the 0 error mark (i.e. visually "less-peaked"). This is equally so demonstrated by the 10% and 90% quantiles, residing farthest from the median. All distributions are observed to be slightly left-skewed, as indicated firstly by shape and secondly by the fact that 10% quantiles are farther removed from the median as compared with the 90% quantiles. Despite that this left-leaning tendency is observable for all network architectures, it is most profound for the PPG and multimodal architectures.

We further explore the results by means of a graphical visualization of absolute window prediction error against label value. This relationship is defined by Figure 5.

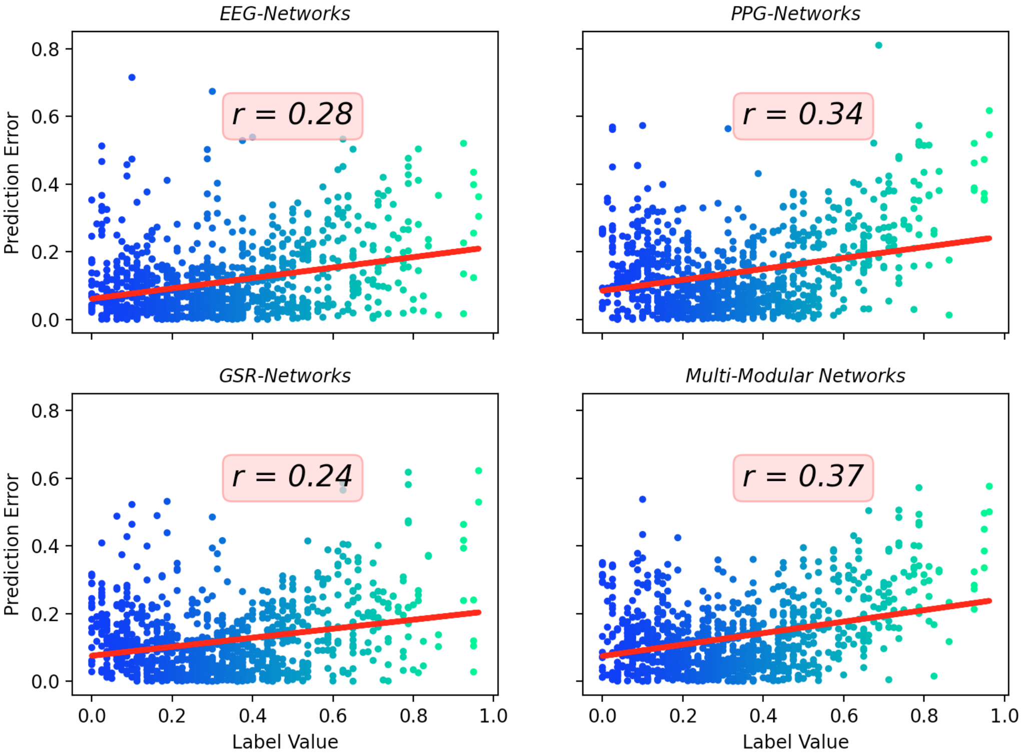

Figure 5. Relationship of absolute window prediction error opposed to window label value. The red lines represent the linear relationship between absolute prediction error and label value. The values r represent the Pearson's Correlation coefficient between prediction errors and label values.

It becomes apparent from Figure 5 that higher labeled windows are substantially less prevailing, as indicated by the thin spreading of dots at the right hand side of the figures, This is conform to the conclusion drawn from the earlier depicted Figure 3. Most noteworthy is that for each of the four architectures, a linear relationship between the height of the absolute window prediction error and its respective label is observed. This is indicated by both the moderately ascending slope and the moderately high (positive) correlations coefficients r. Substantively, this implies that all network architectures are considerably less accurate in predicting high labeled windows. This tendency is strongest for the multimodal architecture. For both the EEG- and GSR -architectures, this linear relationship is observed to be weaker, although still of noteworthy size.

4 Discussion & Conclusion

We found the unimodal EEG architecture to perform best on all performance metrics, with a rather sizable difference in the correlations between predictions and labels. In terms of absolute window prediction error, statistical evidence was found in favor of the unimodal EEG architecture outperforming the multimodel architecture, but not so in favor of the EEG architecture outperforming the GSR architectures. Performance of the GSR architecture was detected to be worse on all metrics as compared with the EEG architecture, but better on all metrics as compared with the multimodal architecture. These differences were, however, not found to be statistically significant in terms of absolute window prediction error. The unimodal PPG architecture was found to perform worst across the board. In terms of absolute window prediction error, this difference was found to be statistically significant when evaluated against the EEG- and GSR -architectures. When applied on the training data, network performance was substantially improved, indicative of overfitting. This especially held true for the singlemodal architectures. Lastly, a moderate relationship was observed between absolute prediction error and label size, implying that all network architectures performed worse on higher labeled windows. This deviation was strongest for the two worst performing architectures, i.e. the multimodal and PPG.

4.1 Discussion

Perhaps the most noteworthy of all aforementioned results is that the unimodel EEG architecture outperformed the multimodal architecture. As argued in the introduction, the opposite was expected: given that the predictions of a multimodal architecture rest on a richer conglomerate of information, better performance is expected. Lending credence to this this inclination is previous research that found a multimodal architecture to perform superiourly (Dolmans et al., 2020; Han et al., 2020; Rastgoo et al., 2019; Yin et al., 2017). In the subsequent sections we aim to rationalize our results.

One reason could be the network architecture. Conform to Han et al. (2020), a shallow head network architecture was adopted, solely consisting of several fully connected dense layers. Further investigation into various alternative head network architectures could have resulted in a more adequately performing network, but would have increased the computational expenses considerably.

Besides this, an important consideration in the fields of statistics and machine-learning alike, is the oftentimes precarious balance between model complexity and simplicity. A model that is overly complex is in increased peril of overfitting, i.e. performing well on training data, but failing to generalize to testing data (Lever, Krzywinski, & Altman, 2016). Overfitting can especially be anticipated when training a relatively complex DNN on a small amount of data (Feng, Zhou, & Dong, 2019; Aggarwal, 2018). As was argued before, our case required personally trained networks due to the person-specific nature that is inherent to physiological data, resulting in a small amount of training data per DNN. Strong evidence in favor of overfitting was found for the unimodal architectures, posing that these network architectures were too complex to be trained on a small amount of data. Surprisingly, the even more complex multiodal architecture did only produce moderate overfitting. However despite this, it should be acknowledged that the core of the multimodal architecture is a composite of the overly complex unimodal architectures, which likely contributed to its relatively inadequate performance.

Another noteworthy result was the expounded relationship between a window's absolute error and its respective label score (as was visualized by Figure 5). It was visualized that all four architectures were consistently less accurate in predicting high labeled windows. A credible explanation for this is the relatively minor presence of high labeled windows in our data (as was visualized by Figure 3). Naturally, when a relatively small amount of high labeled windows are present in the training data, the network doesn't adequately learn to predict such windows.

Finally, despite that a comparison of absolute window prediction error across several architectures was found to be statistically significant, all differences in terms of MAE and RMSE were rather small. The statistical significance of some of the conducted t-tests can particularly be attributed to the vast amount of degrees of freedom with which was tested. We therefore reflect on the following: in opposite to our finding, suppose that a multimodal approach would have entailed a moderate improvement in performance over an unimodal approach. Would this justify the addition in complexity, the considerable increment in computational expenditure and the necessity to invest in a multitude of different physiological sensors? In other words, isn't the already adequately performing, and much less complex unimodal EEG approach sufficient for most tasks at hand? The answer to this question is context dependent, and should be reconsidered in each case specifically.

4.2 Limitations & Future Research Recommendations

Additional research is necessary in order to find more optimal DNN architectures for small data scenario's such as the current. More specifically, 1) which network architectures are able to minimize overfitting and 2) which network architecture is most suitable for a multimodal approach. An effective approach to investigate this is by means of HPO. Apart from only optimizing hyperparamaters, a network architecture in its entirety can be optimized. In the current study we optimized several hyperparamaters, however the architecture in itself was pre-specified. A recommendation for future research is therefore to further investigate upon an appropriate architecture for both the unimodal and multimodal approaches

As expounded on before, all architectures were found to be less precise in predicting high labeled windows, attributable to the fact that such windows were relatively uncommon in our data. This could stem from either 1) the fact that respondents refrained from indicating high workload on the NASA-TLX items utilized for labeling, or 2) respondents simply did not experience high workload throughout the experiment. Alternative labeling approaches may be suited to address the first point. Ideally, one would want to contrast various labeling strategies with one another, such that the most suitable labels for the task at hand may be utilized. Potentially interesting alternatives approaches to labeling are for example a standardized version of the NASA-TLX, or an objective difficulty score for each task, as was proposed by Dolmans et al. (2020). The second point could be addressed by designing the simulational scenario such high workload envoking segments are prevalent throughout the entire scenario. One could assess this by pretesting, and increasing the difficulty in certain segments if need be.

Finally, as was elaborated on before, the ultimate goal of this line of research is to obtain trained networks utilizable for workload assessment in a simulational setting. The personal differences that are inherent to physiological data were acknowledged by utilizing an individual network for each participant. However apart from such between-personal differences, physiological data might be characterized by between-occasional differences as well. To give an example, PPG measurement is dependent upon various occasional factors such as amongst others the amount of emitted surrounding light, body position, cardiovascular stresses, etc. (Lemay et al., 2014). Further research is required in order to assess the generelizability of a personally trained network to different occasions.

4.3 Conclusion

We took three unimodal and one multimodal approach to workload prediction by means of physiological sensors. The performance of the unimodal EEG-, GSR-, and PPG -architectures were assessed in order to determine the most suitable modality for workload prediction. In addition, an architecture combining all aforementioned modalities was proposed and assessed in terms of performance. It was found that all architectures performed rather adequately. A difference in performance was observable however: most notably it was found that the unimodal EEG architecture performed best. Surprisingly, we found a multimodal approach to be unjustified, however several explanations were considered that could potentially refute this. The implications of the current research is another step in the exploration of multimodal deep learning approaches towards workload prediction. Future research in this field will ideally investigate upon 1) network architectures that prevent substantial overfitting in a small dataset scenario, 2) an appropriate architecture for a multimodal approach in a small dataset scenario, 3) the feasibility of different labeling strategies and 4) the cross-occasional generelization of individually trained networks.

References

Aggarwal, C. (2018). *Neural networks and deep learning* (Vol. 10). Springer.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. , 2623–2631. https://doi.org/10.1145/3292500.3330701

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, *13*(1), 281–305.

Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., … Stibler, K. (2005). Evaluation of an eeg workload model in an aegis simulation environment. , *5797*, 90–99. https://doi.org/10.1117/12.598555

Biswas, D., Everson, L., Liu, M., Panwar, M., Verhoef, B., Patki, S., … Helleputte, N. (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. https://doi.org/10.1109/TBCAS.2019.2892297

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. https://doi.org/10.1088/1741-2552/ab0ab5

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. (2020). Percieved mental workload detection using intermediate fusion multi-modal networks. manuscript submitted for publication. https://doi.org/10.3389/fnhum.2020.609096

Feng, S., Zhou, H., & Dong, H. (2019). Using deep neural network with small dataset to predict material defects. *Materials & Design*, *162*, 300–310. https://doi.org/10.1016/j.matdes.2018.11.060

Gunawardana, A., & Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, *10*(12).

Han, S.-Y., Kwak, N.-S., Oh, T., & Lee, S.-W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324–336. https://doi.org/10.1016/j.bbe.2019.12.002

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. , *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hogervorst, M. A., Brouwer, A.-M., & Van Erp, J. B. (2014). Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in neuroscience*, *8*, 322. https://doi.org/10.3389/ fnins.2014.00322

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

Jimenez-Molina, A., Retamal, C., & Lira, H. (2018). Using psychophysiological sensors to assess mental workload during web browsing. *Sensors*, *18*(2), 458. https://doi.org/10.3390/s18020458

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. https://doi.org/10.1038/nature14539

Lemay, M., Bertschi, M., Sola, J., Renevey, P., Parak, J., & Korhonen, I. (2014). Application of optical heart rate monitoring. , 105–129. https://doi.org/10.1016/B978-0-12-418662-0.00023-4

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: model selection and overfitting.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. , 420–423. https://doi.org/10.1145/2414536.2414602

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., … Lerer, A. (2017). Automatic differentiation in pytorch.

Pretorius, A., & Cilliers, P. (2007). Development of a mental workload index: A systems approach. *Ergonomics*, *50*(9), 1503–1515. https://doi.org/10.1080/00140130701379055

Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108. https://doi.org/10.1109/MSP.2017.2738401

Rastgoo, M. N., Nakisa, B., Maire, F., Rakotonirainy, A., & Chandran, V. (2019). Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications*, *138*, 112793. https://doi.org/10.1016/j.eswa.2019.07.010

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. https://doi.org/10.1002/hbm.23730

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. https://doi.org/10.1016/j.neuroscience.2017.07.048

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Sun, X., Hong, T., Li, C., & Ren, F. (2019). Hybrid spatiotemporal models for sentiment classification via galvanic skin response. *Neurocomputing*, *358*, 385–400. https://doi.org/10.1016/j.neucom.2019.05.061

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.

Yin, Z., Zhao, M., Wang, Y., Yang, J., & Zhang, J. (2017). Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, *140*, 93–110. https://doi.org/10.1016/j.cmpb.2016.12.005

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. https://doi.org/10.1080/00140139.2014.956151

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. https://doi.org/10.1080/10447318.2018.1461763

Zhou, J., Jung, J. Y., & Chen, F. (2015). Dynamic workload adjustments in human-machine systems based on gsr features. , 550–558.

A Appendix

Performance on Training Data

Table A.1. Deep neural network performance metrics on training data

	MAE (sd) 0-1 scale	MAE (sd) 0-20 scale	RMSE (sd) 0-1 Scale	Correlation
EEG	0.038 (0.06)	0.761 (1.14)	0.070	0.944
PPG	0.077 (0.11)	1.541 (2.28)	0.114	0.838
GSR	0.085 (0.12)	1.700 (2.43)	0.122	0.813
Multimodal	0.116 (0.16)	2.331 (3.11)	0.156	0.670

Values between parentheses are standard deviations. Mae refers to Mean Average Error. RMSE refers to Root Mean Square Error.