Selecting Feature Sets and Comparing Classification Methods for Cognitive State Estimation

Kati Pettersson
Technical Research Center of Finland
Espoo, Finland
kati.pettersson@vtt.fi

Jaakko Tervonen
Technical Research Center of Finland
Oulu, Finland
jaakko.tervonen@vtt.fi

Johanna Närväinen
Technical Research Center of Finland
Kuopio, Finland
johanna.narvainen@vtt.fi

Pentti Henttonen
University of Helsinki
Helsinki, Finland
pentti.henttonen@helsinki.fi

Ilmari Määttänen
University of Helsinki
Helsinki, Finland
ilmari.maattanen@helsinki.fi

Jani Mäntyjärvi
Technical Research Center of Finland
Oulu, Finland
jani.mantyjarvi@vtt.fi

***Abstract*—Acute stress and high workload are part of everyday work at safety critical fields (e.g. health care). Adaptive human computer interaction systems could support and guide a nurse or a doctor in these hectic situations. Seamless interaction between human and computer requires accurate cognitive state estimation of the person. Currently studies are mainly focused on detecting between two cognitive states with full set of physiologically inspired features. This study demonstrates a classification of different types of stress during Maastricht Acute Stress Test by using feature combinations from electro-oculogram (EOG) and electrocardiogram (ECG) signals in general and personalized approaches, comparing three different classifiers. The classification is evaluated for features extracted from both signals separately and together, and the most important features are selected and reported. Results indicate that the best performance is achieved when features from both EOG and ECG signals are used, and approximately twenty features from EOG and ECG signals are enough to distinguish the two/three states. A personalized approach together with feature selection and support vector machine classifier achieves accuracies of 96.9% and 86.3% in classifying between two states (relaxation and stress) and three states (relaxation, psycho-social stress, and physiological stress), respectively, which exceed state-of-the-art performance. Thus cognitive state estimation benefits from combining selected eye and heart parameters which suggests a promising basis for real-time estimation in the future.**

***Index Terms*—classification, cognitive state, feature selection, electrocardiogram, electro-oculogram, machine learning**

I. Introduction

Rapidly evolving wearable sensors open up completely new possibilities for health care sector to monitor humans unobtrusively and robustly in the field, real life. New sensors could be used for monitoring patients remotely, enabling more accurate diagnosis, treatment as well as plans for rehabilitation. Knowledge of the cognitive state of the nurses and medical doctors (e.g. working at the emergency room/unit) could be used to support their decision making in the situation where the cognitive readiness is limited and the likelihood of human error is increased (e.g. acute stress and other load factors). Moreover, unobtrusive monitoring of human cognitive state in real-life could potentially be helpful in human computer interaction (HCI) application areas such as security and transportation [1].

Intense acute stress impairs decision-making and attention resources. However, often important, even crucial decisions must be made under compromised cognitive resources e.g. in emergency medicine and rescue missions (see review by Starcke & Brand [2]). Self-awareness of the cognitive state is crucial for competent performance, and automation in HCI could help to recognize the status and give advice in critical situations. This would require seamless operation between the HCI system and the human, and the cognitive state estimation should be accurate.

Brain activity measured with electroencephalography (EEG) is a commonly used method to estimate the human cognitive state in HCI (e.g. [3]–[5]). The advances in EEG technology enable wireless monitoring with reasonable signal-to-noise ratio but as EEG is inherently sensitive to different kinds of real-life artifacts such as movement, chewing, blinks etc., the use of EEG data in HCI is challenging. Typically EEG-based classification setups are based on multi-channel measurements leading to several of input variables, long processing time as well as statistical cost (e.g. [3]).

The changes in the cognitive state are reflected also in unconscious behavior and various biosignals. The rapid development of wearable technologies (e.g. heart rate monitoring) and especially smart eyewear glasses has created new platforms for monitoring human cognitive state outside the laboratory more unobtrusively than before. Such measurements with advanced data-analysis techniques provide an opportunity to estimate and further to predict changes in the cognitive state of a person in real-life settings.

Traditionally autonomous nervous system (ANS) signals, e.g. heart rate and its variability (HR, HRV) and skin conductance, are used to assess acute stress and cognitive load (e.g. [6]). The sensitivity and especially the accuracy of biosignal -based stress detection in real-life settings is rather poor and would benefit from complementary information. It has been suggested that mental activities may affect the eye

2471-7819/20/$31.00 ©2020 IEEE
DOI 10.1109/BIBE50027.2020.00115

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.

dynamics: blink duration, blink rate as well as blink rate variability and its complexity [7], [8]. Also the number of saccades (eye movements) has been shown to decrease with increasing heart rate in cognitively stressful situations [9]. Saccades per second [3] and saccade amplitude [10] have been used successfully to discriminate different cognitive states. In [3], seven eye metrics, measured with two small cameras mounted on a lightweight headband, were used to discriminate between cognitive states based on task condition using neural networks (NN) and discriminant function (DF) with different parameters. The classification rates ranged from 69% to 92% depending on the method and task across the study (the best results: Alert v. fatigue (NN): 92%; focused vs. distracted (DF): 85%; relaxed vs. engaged (DF): 87%). In addition, the physiological response to stressor tasks differs between individuals [11], which indicates that taking the individual differences into account, i.e. personalizing the model, is likely to improve its performance.

The performance of different classifiers depend on the employed features, and selection of most important features may help to improve classification results and to identify a minimal feature set to achieve desired classification accuracy. However, in cognitive state detection feature selection is often overlooked and a full set of features available is used. When feature selection has been considered, it has been done by comparing which signal (ECG, skin conductance, respiration, etc.) provides the best set of features for classification [12] or if features across the different signals have been selected, it has not been reported which features were the most important [13], [14]. Selected features may also be extracted from a single signal [15] even though multimodal (employing several signals describing activity of ANS) approaches reach better performance [16]. Moreover, earlier studies including eye parameters in classification have used only a small number of features, e.g. seven features in [3], and eight features in [10], and full potential of different eye parameters has not been utilized.

Maastricht Acute Stress Test (MAST) [17] is a subjective stress-eliciting task, developed to quickly and effectively activate the human stress response. The test consists of alternating trials of physical pain/discomfort (immersion of hand in cold water) and psycho-social stress (mental arithmetic task with time pressure and penalization). It has been suggested that these two stressors have significantly different impact on the stress response induced by MAST [17]. As whole, MAST increases the blood pressure and salivary cortisol levels [17], [18]. However, in a recent study, the average of heart rate (HR, beats per minute) in baseline, MAST, and recovery phases was not a very sensitive indicator of the cognitive state [18]. This limited HR reactivity could imply that the within-MAST HR changes, rise in the arithmetic and drop in the cold-pressor trials, may have cancelled each other in the averaging. As MAST induces a sequence of two different types of stress, it would be desirable to threat these phases as two different cognitive states rather than as a single block of general stress.

Taken together, the state-of-the-art in cognitive state estimation focuses on classification between two stages (binary classification). The aim in this study is to bring the cognitive state estimation closer to the real-life by classifying cognitive states that are (psychophysiologically) close to each other - physiological stress and psycho-social stress. Since the stress reaction is individual, we investigate whether adaptive algorithms for personalizing improve the classification performance. Further, our aim is to improve the classification by including a large set of EOG features and combining them with ECG features, and by finding the most important features for classification.

II. Materials and Methods

A. Participants

Healthy young adults (N = 24, 7 male) volunteered for the study, with mean age of 23.5 years (sd = 3.0 years, range 19 – 29 years). The recruitment criteria were: right-handed, no history in cardiac disorders, no severe depression and no consumption of ANS-affecting medicines.

B. Measurements

The MAST was executed as a part of a larger measurement protocol including four tasks (see Fig 1.) executed in a randomized order. Before every task, a 120s task-baseline as well as a 60s task-anticipation were executed to ensure that the participant had recovered from the previous task; these were not analysed in this paper. Instead, the averages of responses in BL1 and BL2 were used as a baseline for the MAST responses. During BL1 and BL2, the participants were instructed to stay still and keep their eyes on the fixation point on the screen. In MAST the participant alternates between cold-pressor trials (cold immersion of hand in water held at a constant temperature of 2°C (*mastcold*) with varying duration (ranging from 45 – 90s) and mental arithmetics. During the mental arithmetics task (*mastmath*), the participant was asked to verbally perform fast and accurate subtractions: count down from 2043 in steps of 17 under time pressure and facing penalization for mistakes; i.e. they go back to 2043. During both *mastmath* and *mastcold* condition the instructions were visible on the computer screen and in the baseline the participants were instructed to look at a fixation point located in the middle of the computer screen.

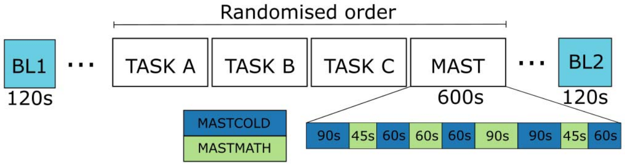

Fig. 1. The study protocol. Only the data from baselines (BL1 and BL2), and MAST are discussed in this work.

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.

C. Signal processing and feature extraction

The psychophysiological signals were measured with NeurOne system (Bittium, Oulu, Finland). One-lead electrocardiogram (ECG) was measured between left collarbone and right lower back. Electro-oculogram (EOG) was measured between the electrodes placed above and beneath the left eye (vertical) and the outer corners of the eyes (horizontal). The signals were sampled at 1000Hz and low-pass filtered with 250Hz cut-off frequency. The inter-beat intervals (IBIs) were extracted from ECG data using an open Matlab toolbox employing Pan-Tompkins algorithm [19]. The saccades and blinks were extracted from the EOG signal with an automated algorithm [20]. Both ECG and EOG features were extracted in 45 second windows with 15 second overlap. The window length was set to maximal that was allowed by the protocol since many heart rate variability (HRV) features require a longer window [21].

The extracted features are listed in Table I. On the ECG signal, statistical features of IBIs and HR were extracted, as well as HRV in time, frequency, geometrical, and non-linear domain. ECG features were extracted with an open-source Python library *hrv-analysis* [22]; for descriptions of each feature we refer to [21] and [23]. On the EOG signal, statistical features as listed in Table I were extracted for saccade rate (SR), time between saccades (TBS), saccade duration (SDUR), blink rate (BR), time between blinks (TBB), blink duration (BDUR), and blink waveform skewness (eye closing time/blink duration, BSKEW), together with their baseline corrected versions. Baseline correction was executed individually by subtracting the median of the baseline measurements (calculated over both BL01 and BL02, see Fig. 1) from each value.

The cognitive state estimation was modelled as a three-class classification problem, with classes corresponding to baseline, *mastcold* and *mastmath*, and a binary problem, which combined the *mastcold* and *mastmath* classes.

D. Mixed Models

When estimating the cognitive state of a person it is important to define the state of interest - the ground truth [24]. The MAST is a validated stress test and it has been reported to elicit physiological stress reactions [17], [18]. To see whether this holds for the collected dataset, a linear mixed model analysis was applied to assess whether the protocol elicited different cognitive states (rest = baseline, physiological stress = *mastcold*, psycho-social stress = *mastmath*), using parameters that have been found to reflect stress: heart, saccade, and blink rate parameters (HRmean, HRstd, SRmean, SRstd, BRmean, BRstd), and root mean square of successive differences of heartbeats, saccades, and blinks (RMSSD, TBSrmssd, TBBrmssd); see Table I for descriptions of parameters. In the mixed model analysis, task type was used as a fixed effect and participant as a random effect. The analysis was divided into two phases: in the first phase, baseline and MAST were used as different task types and in the second phase the MAST condition was further separated into *mastcold* and *mastmath*. The models were implemented by using R's lmer function from the lme4 package [25]. Satterthwaite's method was used to compare models with t-tests and in the second phase the post-hoc analyses were executed by using Tukey's contrasts.

TABLE I
EXTRACTED FEATURES

Signal	Feature	Description
ECG	IBImean, IBImed, IBIrg	IBI mean, median, range
	HRmean, HRstd	HR mean, std
	HRmax, HRmin	HR maximum, minimum
	SDNN, SDSD	Std of IBIs and successive differences
	(p)NN20, (p)NN50	Percentage and number of IBIs differing more than 20ms/50ms
	RMSSD	Root mean square of successive differences
	CVNNI, CVSD	Ratio of SDNN and mean IBI, and RMSSD and mean IBI
	VLF, LF, HF, TotPow	Power in very low, low, high frequency bands, and total power
	LF/HF	Ratio of LF and HF
	LFNU, HFNU	Normalised LF and HF
	HRVTI	HRV triangular index
	CVI, CSI, modified CSI	(modified) cardiac sympathetic index, cardiac vagal index
	SD1, SD2, SD2/SD1	Poincaré plot std perpendicular and along the identity line, their ratio
EOG	X ∈ {SR, TBS, SDUR, BR, TBB, BDUR, BSKEW}, Xmean(BLC), Xmed(BLC), Xstd(BLC), Xrmssd(BLC), Xcv(BLC), Xkurt(BLC), Xskew(BLC), Xp5(BLC), Xp25(BLC), Xp75(BLC), Xp95(BLC)	For saccade rate (SR), time between saccades (TBS), saccade duration (SDUR), blink rate (BR), time between blinks (TBB), blink duration (BDUR), and blink waveform skewness (BSKEW), mean, median, std, rmssd, coefficient of variation, kurtosis, skewness, and 5th, 25th, 75th, and 95th percentiles, and baseline corrected version of each (BLC)

E. Classification Models

Three classification algorithms were applied and compared within the analysis:

- Support Vector Machine (SVM) finds optimal hyperplanes to separate the data. The data are transformed into a dimension in which the classes are linearly separable [26]. Two transformations were employed in this study: linear and radial basis function (RBF).
- Random Forest (RF) is an ensemble of decision trees. Each decision tree separates the data in a hierarchical fashion, aiming to contain only data originating from a single class in each leaf node. In a RF, the trees are built independently using a random sample of data and features for each tree [27].
- XGBoost (XGB) is another type of ensemble of decision trees. However, the trees in XGB are not independent but each subsequent tree attempts to fix the errors made by preceding trees [28].

The scikit-learn implementation of SVM and RF [29] and the Python implementation of XGB [28] were used. To achieve the best classification accuracy the hyperparameter configuration for each model was chosen heuristically by fitting each

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.

model with different hyperparameters and choosing the most accurate for final estimation. Due to space restrictions, the final hyperparameters are not reported here but are available from the authors on request.

F. Model Evaluation and Performance Measures

A non-personalized and a personalized model were trained with each ML method described. The adopted personalization method was within-subject normalization: the feature values were normalized by subtracting the mean and dividing by standard deviation for each subject separately. An eight-fold cross-validation was used so that the data of three randomly selected subjects was left out at a time from training (a leave-three-subjects-out validation) and performance was measured on the left out subjects. Accuracy and F1-score were considered as measures of performance. Accuracy is the ratio of correctly classified samples. F1-score is the harmonic mean of recall, probability of detecting each class, and precision, reliability of results in each class. The final F1-score was obtained by calculating recall and precision separately for each class and averaging them, weighted by the number of samples in each class.

G. Determining Most Important Features

The Sequential Forward Floating Search (SFFS) [30] was used together with SVM to find the most important features. SFFS is an algorithm that alternates between two steps: inclusion and exclusion. At each step, the algorithm finds one feature to add to selected features (i.e. includes the one maximizing the relevant criterion function). After inclusion, the algorithm checks whether excluding a selected feature improves the criterion function; if so, the feature is excluded and another exclusion step is executed. If no feature exclusion improves the criterion function, the algorithm returns to inclusion step. The SFFS implementation in MLxtend Python library [31] was adopted.

SFFS was not executed together with the tree-based models because it was computationally infeasible. Instead, the importance of each feature for RF and XGB was assessed in terms of relative impurity reduction: the importance is the normalized total reduction of impurity brought by that feature [28], [29]. The Gini index was used as a measure of node impurity.

As classification results (see Section III-B) showed that personalized models with both ECG and EOG features was the most accurate, feature importance estimation was conducted only for this condition.

III. Results

A. Mixed Model Results

Linear mixed models showed statistically significant difference in binary comparison (between baseline and MAST) and in the three class comparison (between baseline, *mastmath* and *mastcold*) for all the selected heartbeat (see Fig. 2) and blink, and for some of the saccade features (see Table II). SRmean did not differ statistically significantly between the baseline and *mastcold*, whereas SRstd did not differ between *mastcold* and *mastmath*. In both binary and three-class comparison the task type did not explain the variation in TBSrmssd -feature and the model did not differ statistically from the null model, which is why the test statistics and p-value could not be reported for the feature.

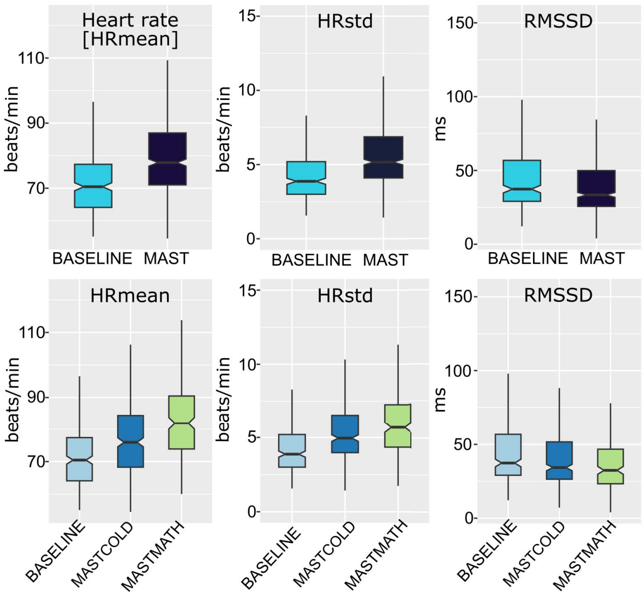

Fig. 2. Whiskers boxplots of HRmean, HRstd and RMSSD.

B. Classification Results

The classification performance on the three-class and binary tasks are reported in Tables III and IV, respectively.

In the three-class task, personalized models always performed better than non-personalized. Eye-related features performed better than ECG features but the highest scores were obtained when both were used. XGB was generally the most accurate model, achieving the highest accuracy of 84.5% with a personalized model when both types of features were used. Table V shows the confusion matrix of best performing three-class classifier: Nearly two thirds of misclassifications were between *mastcold* and *mastmath*.

Also in the binary classification, personalization proved useful but in some conditions (when both ECG and EOG features were used together with RF or XGB) the non-personalized version performed equally. The best accuracies with a personalized and a non-personalized model were 94.4% and 94%, respectively, with RF classifier.

C. Feature Importance Results

The SFFS algorithm was run for personalized three-class and binary classification with SVM, using linear and RBF transformations. The progress of the algorithm is shown in Fig. 3. The accuracies started at around 68% and 85% for three-class and binary problems, respectively, and steeply increased to around 85% and 95% when ten features were selected. As more than ten features were added, the accuracy

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.

TABLE II
MIXED MODEL RESULTS.

Parameter	Binary comparison		Three-way comparison					
	MAST – BL		COLD - BL		MATH - BL		MATH - COLD	
	T-value	P-value	Z-value	P-value	Z-value	P-value	Z- value	P-value
HRmean	18.82	< 0.001	13.38	< 0.001	25.55	< 0.001	14.44	< 0.001
HRstd	10.58	< 0.001	8.38	< 0.001	10.39	< 0.001	3.26	0.003
RMSSD	-6.45	< 0.001	-3.91	< 0.001	-8.27	< 0.001	-5.04	< 0.001
BRmean	15.33	< 0.001	11.13	< 0.001	18.17	< 0.001	8.92	< 0.001
BRstd	12.3	< 0.001	8.45	< 0.001	15.26	< 0.001	8.33	< 0.001
TBBrmssd	-7.47	< 0.001	-6.25	< 0.001	-6.87	< 0.001	-1.54	0.271
SRmean	2.51	0.012	1.24	0.4284	3.61	< 0.001	2.64	0.023
SRstd	4.91	< 0.001	3.69	< 0.001	5.10	< 0.001	1.99	0.115
TBSrmssd	-	-	-	-	-	-	-	-

Abbreviations: BL: baseline, COLD: *mastcold*, MAST: *mastcold* and *mastmath* together, MATH: *mastmath*

TABLE III
CLASSIFICATION RESULTS ON THE THREE-CLASS (BASELINE VS. *mastcold* VS. *mastmath*) CLASSIFICATION TASK.

Features	Personal	Linear SVM		RBF SVM		RF		XGB	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
ECG	No	**56.6 (6.6)**	**53.3 (7.8)**	54.3 (7.9)	50.9 (9.0)	52.9 (7.2)	50.1 (8.8)	50.9 (5.0)	49.2 (5.9)
	Yes	**70.2 (10.6)**	**70.1 (10.5)**	68.0 (10.8)	68.0 (10.7)	68.4 (9.1)	68.3 (9.0)	68.8 (9.6)	68.8 (9.5)
EOG	No	63.6 (7.7)	63.0 (7.0)	67.9 (7.1)	67.8 (7.0)	77.3 (6.6)	77.2 (5.9)	**78.1 (6.8)**	**78.3 (6.2)**
	Yes	75.1 (7.7)	74.6 (7.8)	75.8 (8.1)	75.3 (8.0)	78.4 (7.3)	78.2 (7.3)	**79.9 (8.2)**	**79.5 (8.2)**
ALL	No	65.3 (5.8)	63.7 (6.0)	68.2 (8.5)	67.5 (9.1)	76.5 (9.5)	75.5 (10.5)	**78.7 (8.6)**	**78.5 (8.5)**
	Yes	81.4 (8.9)	81.4 (8.9)	82.5 (8.1)	82.5 (8.0)	83.9 (8.8)	83.7 (9.0)	**84.5 (7.9)**	**84.4 (8.0)**

Values are means with standard deviations in parentheses.

TABLE IV
CLASSIFICATION RESULTS ON THE BINARY (BASELINE VS. MAST) CLASSIFICATION TASK.

Features	Personal	Linear SVM		RBF SVM		RF		XGB	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
ECG	No	**74.1 (7.5)**	**73.1 (8.0)**	72.0 (6.8)	70.0 (7.4)	71.5 (8.0)	69.9 (8.2)	70.7 (6.6)	68.9 (6.9)
	Yes	**83.1 (8.0)**	**82.9 (8.0)**	81.7 (9.7)	81.2 (10.0)	82.9 (7.9)	82.7 (7.9)	82.0 (8.0)	81.6 (8.2)
EOG	No	84.8 (5.5)	84.5 (5.5)	89.9 (5.6)	89.6 (5.7)	**93.8 (2.6)**	**93.6 (2.8)**	93.2 (3.5)	93.0 (3.7)
	Yes	89.3 (3.6)	89.2 (3.6)	90.0 (5.4)	89.9 (5.5)	90.5 (4.8)	90.3 (4.9)	**91.4 (4.3)**	**91.3 (4.4)**
ALL	No	85.9 (5.5)	85.7 (5.6)	90.2 (4.8)	90.1 (4.9)	93.4 (3.4)	93.1 (3.8)	**94.0 (4.0)**	**93.7 (4.5)**
	Yes	92.5 (5.0)	92.4 (5.2)	92.8 (4.3)	92.7 (4.4)	**94.4 (2.8)**	**94.4 (2.8)**	94.0 (2.7)	94.0 (2.7)

Values are means with standard deviations in parentheses.

did not improve much anymore, and maximal accuracies of 86.3% and 96.9% for the three-class and binary problems, respectively, were obtained with 21 and 25 features. As the RBF transformation provided higher accuracy than the linear transformation, the features selected and accuracy obtained at each step of the algorithm are shown in Table VI for both prediction tasks using the RBF SVM model.

The relative feature importances obtained with XGB and RF in the three-class task are shown for the top-30 features in Fig. 4. As the values of relative importances with the RF are higher than with XGB, XGB weighted different features more uniformly but still both models used all available features to do some data splits (all features had strictly positive importance).

TABLE V
CONFUSION MATRIX OF XGB MODEL WITH BOTH ECG AND EOG FEATURES AND PERSONALIZED MODEL.

		Predicted		
		Baseline	Cold	Math
True	Baseline	245	19	3
	Cold	21	280	31
	Math	3	45	143

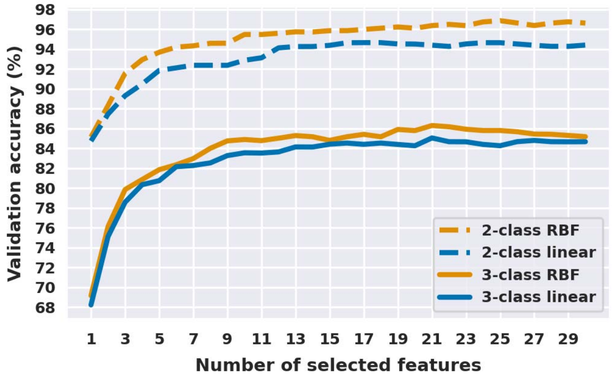

Fig. 3. The accuracy of SVM with linear and RBF kernel during the sequential feature search.

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
FEATURES SELECTED BY THE SFFS ALGORITHM WITH RBF SVM CLASSIFIER UNTIL BEST ACCURACY REACHED.

Feature selected	Three-class RBF SVM		Binary RBF SVM	
	Feature	Accuracy	Feature	Accuracy
1st	HRmean	69.1	SDURp75	85.1
2nd	SDURmed	76.1	TBBp25	88.3
2rd	TBBmed	79.8	Hrmax	91.6
4th	SD2	80.9	BDURsd	92.9
5th	BRp95	81.8	TBScv	93.7
6th	SDURp25	82.3	HF	94.2
7th	BDURmean	83.0	BSKEWp25	94.3
8th	BDURcv	84.0	SDURp75BLC	94.6
9th	SDURp5	84.8	LF/HF	94.6
10th	TBSsd	84.9	SDURsd	95.5
11th	TBSp95	84.8	SDURp95	95.5
12th	SDNN	85.0	Hrmin	95.6
13th	pNNI20	85.3	pNNI20	95.7
14th	TBSmean	85.2	SDNN	95.7
15th	TBBp25	84.8	TBBp25BLC	95.9
16th	SDURcv	85.2	BDURrmssd	95.9
17th	pNNI50	85.4	SDURcv	96.0
18th	SDURp5	85.2	TBBp25BLC	96.1
19th	TBBp25BLC	85.9	BDURcv	96.2
20th	SDURp5BLC	85.8	IBImed	96.1
21st	IBImean	86.3	BDURskew	96.4
22nd			HRstd	96.5
23rd			Hrmean	96.4
24th			CSI	96.7
25th			CVNNI	96.9

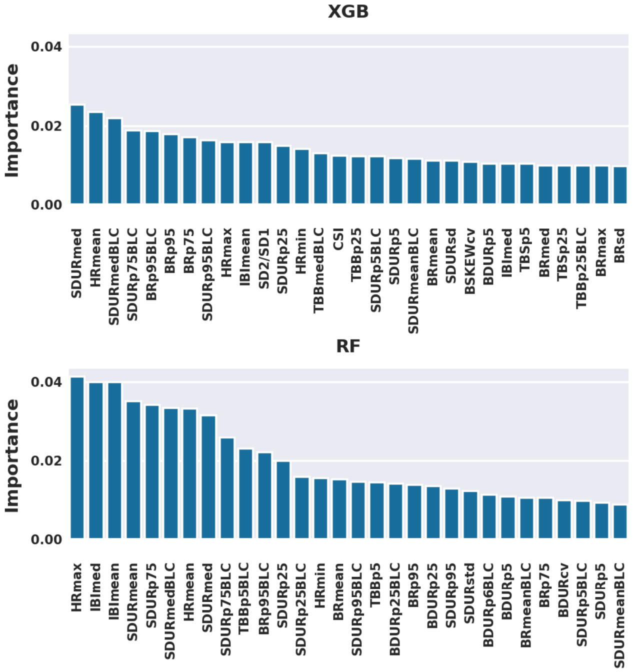

Fig. 4. Relative feature importances of XGB (upper panel) and RF (lower panel) on three-class personalized classification with all features.

IV. DISCUSSION

The objective of this study was to bring the cognitive state estimation closer to the real life conditions, where the cognitive states are not that easily differentiated. The MAST protocol was used to induce two different kind of stress: physiological and psycho-social. Aim was to improve the estimation results between these two stress dimensions by comparing classification methods and feature set selection derived from both EOG and ECG signals.

Linear mixed models were applied to define the cognitive state of interest by using heartbeat, blink, and saccade features that have shown to be sensitive for stress and cognitive load (e.g. [3], [4], [6], [32]). In the binary comparison the MAST increased heart rate and reduced variability. Similar result was seen in the blink features, but the only saccade feature that showed statistical difference between MAST and baseline comparison was the standard deviation of the saccade rate. In the three-class comparison the *mastmath* induced the strongest reactions increasing the heart and blink rates, and decreasing variability in both. Also the *mastcold* increased the heart and blink rates, and decreased variability compared to the baseline. Again, the changes in the saccade parameters were not that consistent.

In earlier literature the increasing heart rate and blink rate as well as decreasing RMSSD (both HR and BR cases) have been reported to reflect increasing stress or cognitive load [8], [32]. Hence, these results imply that the participants experienced the psycho-social stress (*mastmath*) more stressful than physiological stress (*mastcold*) or rest (baseline). Further, the physiological stress was clearly more stressful than the rest condition.

The visual information effects both saccades and blinks, but especially the saccades. In both *mastcold* and *mastmath* the visual stimulus was similar: a few lines of instructions on the screen, whereas in the baseline the participants looked at a fixation point. Therefore, it is consistent that the participants made more saccades during the MAST than in the baseline condition (binary comparison). However, the three class comparison revealed that the participants do not move their eyes statistically significantly more in *mastcold* than in the baseline while in the *mastmath* condition the participants made clearly more eye movements than in other conditions. This could be explained by the task type: in *mastcold* the participants try tolerate the unpleasant, even painful condition and fixating at the same point could help to cope the situation whereas in *mastmath* looking at the starting number and the step number given in the instruction text may help to concentrate better while executing the difficult mental arithmetic task.

Several classifiers were applied to find how well the elicited states can be predicted using ECG, EOG, or both kind of features in a personalized and non-personalized manner (see Tables III and IV). XGB was generally the most accurate model, and personalized models were usually more accurate than non-personalized. However, when both ECG and EOG features were used with RF or XGB classifiers, non-personalized model was basically as accurate as personalized. Thus the two states in the binary case are separate enough that a complex model with a large number of features is able to recognize the states with high accuracy even when

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.

with a non-personalized model. This is also evidenced by the confusion matrix in Table V. As majority of mistakes were made between *mastcold* and *mastmath*, it shows that the two MAST conditions were distinguished from baseline with higher confidence than they were from each other.

When selecting the most important features, ten features with RBF SVM were enough to distinguish the states in both three-class and binary case more accurately than the XGB and RF models using all the features, and most accurate results were found with little more than 20 features for both cases. Although the individual features selected for the two tasks differ, a pattern was observed in the top six features selected in both tasks: the first three features selected contain one feature related to each HR(V), blinks, and saccades, and the next three also contain one feature related to HR(V), blinks, and saccades (see Table VI). The tree-based models showed an almost similar pattern: the most important features in XGB were related to HR, SDUR and BR, whereas in RF they were related to HR, IBI, and SDUR; RF was the only model that did not consider any blink-related feature among the five most important. As the features consisted of three physiological responses (heartbeats, blinks, saccades), the three dimensions seem to provide complementary information on cognitive state and physiology, which is reflected to the feature selection.

The overall highest accuracies obtained in this study were 86.3% and 96.9% in three-class and binary classification, respectively, with personalized RBF SVM model and SFFS feature selection. Previously, classification between different cognitive states has yielded accuracies between 90% and 95% in binary classification [3], [5], [12], [33], [34] and between 73% and 84% in classification with multiple states [10], [12], [13], [35] depending on the states to detect, protocol used to elicit them, and the signals involved and classifier used. Thus, state-of-the-art accuracy was obtained in this study.

Since RBF SVM was more accurate than RF and XGB after feature selection, it shows that a simpler classifier can perform well in cognitive state estimation. However, the computational cost of running the feature selection and simultaneously optimizing SVM parameters is substantial, and therefore it may not be feasible for all applications. SVM also scales poorly for larger datasets, so this approach may be computationally infeasible if there are more subjects, physiological signals or cognitive states involved. Then, RF or XGB may provide a better starting point since they performed better without feature selection and also scale better for larger data.

One difference between the baseline condition and MAST task was the visual information presented to the participants. During the baseline, they were seeing a fixation point in the middle of the screen, while in the MAST task the instructions (two lines of text) were visible. This may have induced systematic differences in eye movements (saccades) during these conditions. In real life, the visual stimulus is extremely difficult to control and therefore the eye movements could be used for classifying different task types (such as reading, visual search etc.) [10], [36] rather than used as a feature reflecting the persons cognitive state directly.

The objective in this study was in detecting cognitive states that are not that easily differentiated. The results showed that the resting condition can be well distinguished from other states but it was more difficult to detect between physiological and psycho-social stress. As the states resemble each other closely and they cause similar physiological reactions, in the future we have plans to model them as continuous rather than distinct, which has already been attempted in the context of emotion recognition [14]. The detection accuracy might also benefit from additional signals, such as skin conductance, respiration or body temperature. Including more signals would bring more insight into finding a minimal set of signals and features for reliable and efficient cognitive state detection.

V. Conclusions

Monitoring human in real life could serve variety of fields to support human cognitive capabilities, recognize and track symptoms as well as plan adequate treatments and medications for the patients. Lots of work needs to be done before the seamless human state estimation is reality, but it can be seen that the potential to improve life in general is huge. This paper takes one step closer to the future; attempting to find the best heart and eye features and method to classify rest, physiological stress, and psycho-social stress from each other. We found that eye parameters classify these three different stress stages better than the heart rate variables alone, and further by combining eye and heart features the classification was even better. However, many of the features were redundant and it was shown that optimal classification performance was obtained with approximately twenty features, including both heart rate and eye-related parameters. As the physiological stress reactions and thus the parameter values are individual, it was shown that personalization improves the classification result. The outcome of this paper is that by careful feature selection and personalization, it is possible to use eye and heart features to classify cognitive states that are physiologically near each other with state-of-the-art performance even with a simple classifier.

Acknowledgment

We would like to thank MSc. Bart van der Sanden for the ECG signal processing and analyses. These measurements were part of SISU Project (Academy of Finland ICT2023 program, decisions: 313401 and 313399).

References

[1] J. van Erp, F. Lotte, and M. Tangermann, “Brain-Computer Interfaces: Beyond Medical Applications,” *Computer*, vol. 45, apr 2012.

[2] K. Starcke and M. Brand, “Decision making under stress: A selective review,” *Neuroscience and Biobehavioral Reviews*, vol. 36, 2012.

[3] S. P. Marshall, “Identifying cognitive state from eye metrics,” *Aviation Space and Environmental Medicine*, vol. 78, 2007.

[4] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, “Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness,” *Neuroscience and Biobehavioral Reviews*, vol. 44, pp. 58–75, 2014.

[5] M. Stikic, R. R. Johnson, V. Tan, and C. Berka, “EEG-based classification of positive and negative affective states,” *Brain-Computer Interfaces*, vol. 1, 2014.

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.

[6] R. J. Croft, C. J. Gonsalvez, J. Gander, L. Lechem, and R. J. Barry, "Differential relations between heart rate and skin conductance, and public speaking anxiety.," *Journal of behavior therapy and experimental psychiatry*, vol. 35, 2004.

[7] T. Abe, T. Nonomura, Y. Komada, S. Asaoka, T. Sasai, A. Ueno, and Y. Inoue, "Detecting deteriorated vigilance using percentage of eyelid closure time during behavioral maintenance of wakefulness tests," *International Journal of Psychophysiology*, vol. 82, 2011.

[8] R. Paprocki and A. Lenskiy, "What does eye-blink rate variability dynamics tell us about cognitive performance?," *Frontiers in Human Neuroscience*, vol. 11, 2017.

[9] F. Dehais, M. Causse, F. Vachon, and S. Tremblay, "Cognitive conflict in human-automation interactions: A psychophysiological study," *Applied Ergonomics*, vol. 43, 2012.

[10] J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk, "Predicting Cognitive State from Eye Movements," *PLoS ONE*, vol. 8, 2013.

[11] K. Oishi, M. Kamimura, T. Nigorikawa, T. Nakamiya, R. E. Williams, and S. M. Horvath, "Individual differences in physiological responses and type A behavior pattern," *Journal of Physiological Anthropology and Applied Human Science*, vol. 18, 1999.

[12] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, (New York, New York, USA), ACM Press, 2018.

[13] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, 2017.

[14] C. A. Torres-Valencia, M. A. Alvarez, and A. A. Orozco-Gutierrez, "Multiple-output support vector machine regression with feature selection for arousal/valence space emotion assessment," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, 2014.

[15] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity," *IEEE Transactions on Affective Computing*, vol. 3045, 2019.

[16] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, 2015.

[17] T. Smeets, S. Cornelisse, C. W. Quaedflieg, T. Meyer, M. Jelicic, and H. Merckelbach, "Introducing the Maastricht Acute Stress Test (MAST): A quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses," *Psychoneuroendocrinology*, vol. 37, 2012.

[18] A. L. Shilton, R. Laycock, and S. G. Crewther, "The Maastricht Acute Stress Test (MAST): Physiological and subjective responses in anticipation, and post-stress," *Frontiers in Psychology*, vol. 8, 2017.

[19] H. Sedghamiz, "Complete Pan-Tompkins Implementation ECG QRS Detector," 2014.

[20] K. Pettersson, S. Jagadeesan, K. Lukander, A. Henelius, E. Hæggström, and K. Müller, "Algorithm for automatic analysis of electro-oculographic data," *BioMedical Engineering Online*, vol. 12, no. 1, 2013.

[21] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, 2017.

[22] R. Champseix, "Heart Rate Variability analysis," 2018.

[23] J. Jeppesen, S. Beniczky, P. Johansen, P. Sidenius, and A. Fuglsang-Frederiksen, "Using Lorenz plot and Cardiac Sympathetic Index of heart rate variability for detecting seizures for patients with epilepsy," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, pp. 4563–4566, 2014.

[24] A. M. Brouwer, T. O. Zander, J. B. van Erp, J. E. Korteling, and A. W. Bronkhorst, "Using neurophysiological signals that reflect cognitive or affective state: Six recommendations to avoid common pitfalls," *Frontiers in Neuroscience*, vol. 9, 2015.

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using {lme4}," *Journal of Statistical Software*, vol. 67, 2015.

[26] C. J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, 1998.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, 2001.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[29] F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, V. Dubourg, and R. Weiss, "Scikit-learn : Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[30] P. Pudil, J. Novovieova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 365–366, 1994.

[31] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack," *Journal of Open Source Software*, vol. 3, 2018.

[32] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, 2015.

[33] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment," *Proc ACM Int Conf Ubiquitous Comput.*, 2015.

[34] O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *International Journal of Neural Systems*, vol. 27, 2017.

[35] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-EEG biosignals dataset for assessment and visualization of neurological status," *IEEE Workshop on Signal Processing Systems, SiPS: Design and Implementation*, 2016.

[36] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Eye movement analysis for activity recognition using electrooculography," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, 2011.

Authorized licensed use limited to: UNIVERSITY OF CONNECTICUT. Downloaded on May 15,2021 at 12:35:03 UTC from IEEE Xplore. Restrictions apply.