Machine-Learning Based Monitoring of Cognitive Workload in Rescue Missions with Drones

Fabio Dell'Agnola, *Member, IEEE,* Ping-Keng Jao, Adriana Arza, *Member, IEEE,* Ricardo Chavarriaga, *Senior Member, IEEE,* José del R. Millán, *Fellow, IEEE,* Dario Floreano, *Senior Member, IEEE* and David Atienza, *Fellow, IEEE*

***Abstract*— In search and rescue missions, drone operations are challenging and cognitively demanding. High levels of cognitive workload can affect rescuers' performance, leading to failure with catastrophic outcomes. To face this problem, we propose a machine learning algorithm for real-time cognitive workload monitoring to understand if a search and rescue operator has to be replaced or if more resources are required. Our multimodal cognitive workload monitoring model combines the information of 25 features extracted from physiological signals, such as respiration, electrocardiogram, photoplethysmogram, and skin temperature, acquired in a noninvasive way. To reduce both subject and day inter-variability of the signals, we explore different feature normalization techniques, and introduce a novel weighted-learning method based on support vector machines suitable for subject-specific optimizations. On an unseen test set acquired from 34 volunteers, our proposed subject-specific model is able to distinguish between low and high cognitive workloads with an average accuracy of 87.3% and 91.2% while controlling a drone simulator using both a traditional controller and a new-generation controller, respectively.**

***Index Terms*— Cognitive Workload Monitoring, Physiological Signals, Machine Learning, Human-Robot Interaction, Wearable Systems, Search and Rescue Missions.**

I. Introduction

THANKS to recent enhancements in both robotics and human-robot interfaces, the interest in deploying robots in search and rescue (SAR) missions is growing [1]. However, limitations exist in their effective and efficient utilization in real-life missions. The main limitation is that robot teleoperation is a non-intuitive and challenging task. Thus, SAR robots are still constrained to simple missions and highly trained professionals. [2], [3]. Moreover, rescuers have to simultaneously focus on multiple tasks and deal with both scarcity of human resources and time pressure. This situation is cognitively highly demanding and can negatively affect performance [4], [5]. Consequently, operating under high cognitive workload (CWL) may severely compromise the execution of a mission and leads to failure with catastrophic outcomes [6]. Therefore, there is a need to monitor CWL to ensure efficient execution of SAR missions.

To assess CWL, researchers typically use surveys [7], performance metrics [8], [9], and information from physiological signals [10]. However, surveys only provide subjective and sporadic measurements, and are not always reliable [11]. Although performance metrics provide objective measurements, reliable metrics are difficult to set as every rescue mission is unique. On the other hand, physiological signals can be noninvasively acquired without disturbing the rescuers' work. Thus, the use of physiological signals seem the most promising solution to assess Cognitive Workload Monitoring (CWM) [10], [12], [13].

Several studies combine physiological signals with different machine-learning algorithms for CWM in different fields [13], [14]. However, to the best of our knowledge, we are the first to address CWM of drone pilots involved in SAR missions [8], [15], [16]. Now, we extend our previous works by presenting a subject-specific CWM approach based on noninvasive physiological signals that is suitable for new drone control solutions, such as FlyJacket [17]. In particular, this work proposes the following contributions:

- We explore different feature normalization techniques to reduce both inter-subject and inter-day variability;
- We provide a new weighted-learning method for Support Vector Machine (SVM), suitable for subject-specific optimizations. This SVM based method uses two regularization terms, one for learning the general behaviour and another for tuning the model to fit the characteristics of a particular data subset;
- We prove the ability of our method to detect low and high CWL levels while controlling a drone simulator with traditional and advanced controllers, achieving an accuracy of 87.3% and 91.2%, respectively. These results are obtained on unseen data acquired from 34 participants while flying a drone simulator and mapping a graphic representation of a disaster situation. Our results are higher than the latest state-of-the-art studies in SAR missions with drones (see Table I).

II. Related Work

CWL characterization and estimation have been addressed by a large number of studies [12], [25], which characterize

This work was partially supported as a part of NCCR Robotics, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (Grant No. 51NF40_185543) and by the ONR-G through the Award Grant No. N62909-17-1-2006.

F. Dell'Agnola, A. Arza, and D. Atienza are with the Embedded Systems Laboratory (ESL), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland (e-mail: fabio.dellagnola@alumni.epfl.ch, adriana.arza@epfl.ch, david.atienza@epfl.ch).

P.-K. Jao, R. Chavarriaga, and J.d.R. Millán are with EPFL, Lausanne 1015, Switzerland. J.d.R. Millán is also with the Dept. of Electrical and Computer Engineering & the Dept. of Neurology, The University of Texas at Austin, Austin, TX 78712, US (e-mail: ping-keng.jao@alumni.epfl.ch, ricardo.chavarriaga@alumni.epfl.ch, jose.millan@epfl.ch).

D. Floreano is with the Laboratory of Intelligent Systems (LIS), EPFL, Lausanne 1015, Switzerland (e-mail: dario.floreano@epfl.ch).

TABLE I
SUMMARY OF THE STATE-OF-THE-ART STUDIES USING MULTIPLE PHYSIOLOGICAL SIGNALS

Study	Performed Tasks	Physiological Signals	Window Length (Overlap)	Classifier (Classes)	Results Acc.	Sens.	Spec.
Momeni et al. [15]	Simulated SAR with drones	ECG, RSP, PPG, SKT	60s (30s)	XGB (2)	86%*	-	-
Dell'Ágnola et al. [18]	Simulated SAR with drones	ECG, RSP, PPG, SKT	60s (0s)	XGB (2)	80.2%*	79.6%*	71.7%*
Montesinos et al. [19]	Arithmetic tasks	ECG, PPG, RSP, SKT, EDA	60s (30s)	RF (2)	84.13%*	-	-
Chen et al. [20]	Real car driving	ECG, RSP, EDA	100s (90s)	SVM (3)	89.7%	88.5%	94.2%
Solovey et al. [21]	Driving in highway	ECG, EDA	30s (0s)	LR (2)	90%	-	-
Giakoumis et al. [22]	Video-game	ECG, EDA	25s (0s)	LDA (2)	94.96%	94.96%	94.96%
Tjolleng et al. [23]	Simulated driving task	ECG	100s (0s)	ANN (3)	82%	78%	91%
Gjoreski et al. [24]	Daily life activities	PPG, SKT, EDA	300s (150s)	SVM (2)	98.96%	70.44%	99.88%

ECG-electrocardiogram, RSP-respiratory activity, PPG-photoplethysmogram, SKT-skin temperature, EDA-electrodermal activity. XGB-Extreme Gradient Boosting, RF-Random Forest, SVM-Support Vector Machine, LR-Logistic Regression, LDA-Linear Discriminant Analysis, ANN-Artificial Neural Network.
* Results based on an unseen test set, all the other are limited to cross-validation.

either the performance or the distress of a person involved in a particular task or situation. In this section, we review the state-of-the-art machine learning (ML) techniques detecting CWL induced by high cognitive tasks. In particular, we analyze those works using unobtrusively measured physiological signals. Although interesting for their results, studies relying on obtrusive measurements (e.g., electroencephalography [26]) are not included in this analysis since their integration into a jacket is difficult or unattainable. The same applies to works placing sensors in locations other than the torso, such as the head [27].

Table I summarizes the most recent and significant studies including the performed task to induce CWL, measured physiological signals, signal segmentation (i.e., window length and overlap), applied machine-learning methods, targeted classes, and classification results (i.e., Accuracy, Sensitivity, and Specificity). Our analysis identifies the following common methodological steps: signal acquisition and preprocessing (filtering and segmentation), feature extraction, feature normalization, dimension reduction or feature selection, and classification or regression. However, although the methodology is well established, discrepancies are found in different steps. Hence, in the following, we review these discrepancies.

First, significant differences have been observed on the physiological measures, which are electrodermal activity (EDA) [19]–[22], [24], [28], electrocardiogram (ECG) [18], [20]–[23], [29], photoplethysmogram (PPG) [15], respiratory activity (RSP) [15], [20], and peripheral skin temperature (SKT) [15], [18]. Although using multiple physiological signals can increase the detection accuracy of CWL levels [15], the type and number of signals, and in particular the features set, often differ and strictly depend on the case study (e.g., the type of task used to induce different levels of CWL) [10], [29]. Thus, there is no clear definition of the best selection of signals and features to assess CWL in general.

Then, the segmentation window used to extract the features from the signals also depends on the case study. In particular, the window lengths reported in Table I vary from 25 to 300 seconds. Moreover, different window overlaps are applied either to increase the size of the dataset [15] or to provide more frequent estimations in time [20], [24]. These differences can be explained by the fact that physiological methods do not provide a direct measurement of the workload, but rather they give information about how the individuals themselves respond to a particular load [10]. So, a different signal segmentation may be applied depending on the dynamic of the physiological response induced by a particular CWL.

An additional aspect observed in our literature review is that features are often normalized to standardize their ranges. The normalization help to reduce intra- and inter-subject variability caused by age, time of day and other factors [30]. However, not all studies report whether a normalization was applied [30], or clearly explaining how it was done and distinguishing between training and test sets. To properly emulate and test the system's behaviour, test data should be normalized based on the parameters obtained from the training set [30].

Moreover, the choice of machine-learning methods clearly differ. The train data size and the system requirements specification (e.g., computational complexity, power and latency) may explain the different selections of machine-learning algorithms. In fact, as most of the studies typically start with a limited amount of data, simple models like Support Vector Machine (SVM) [20], [24], [31], Linear Discriminant Analysis (LDA) [22], [28], Logistic Regression (LR) [21], and Decision Tree (DT) [19], are the most used machine-learning techniques. In contrast, complex models such as Artificial Neural Networks (ANN) [23], Random Forest (RF) [19], [31], and very recent models like Extreme Gradient Boosting (XGB) [15], have been less used so far. In any case, even if SVM has been the most used classifier in this field, there is no consistent indication of whether it is the best model or not for different case studies.

Finally, our review shows that the highest accuracy levels are in the range of 80 to 99%. This wide range is mainly due to the diverse experimental protocols, methodologies, and number of considered classes in each study. Also, the highest accuracies reported by different studies may be affected by overfitting since their model evaluation is limited to cross-validation [20]–[24]. However, a proper estimation of a model's generalization power requires a final test on new unseen data, a set never used in training [15], [18], [19].

In conclusion, there is a need to investigate further the contribution of each physiological signal, the impact of data normalization, and the performance of the selected classifier on unseen data in the context of rescue missions with drones, which are not appropriately covered in the literature.

Besides, workload is multidimensional [7] and results from the aggregation of three broad aspects [10], [32]. First, the workload depends on the task's type (mental or physical demand), and the load level (e.g., tasks amount and difficulty). Second, it is affected by time, namely, by the duration of the temporal demand. Third, the subjective psychological experiences modulate the level of workload perceived by a subject (i.e., subject's capabilities, learning skills, and effort). So, it is necessary to investigate CWL in the particular field of interest and, also, consider each person's subjective workload level, as suggested in [33].

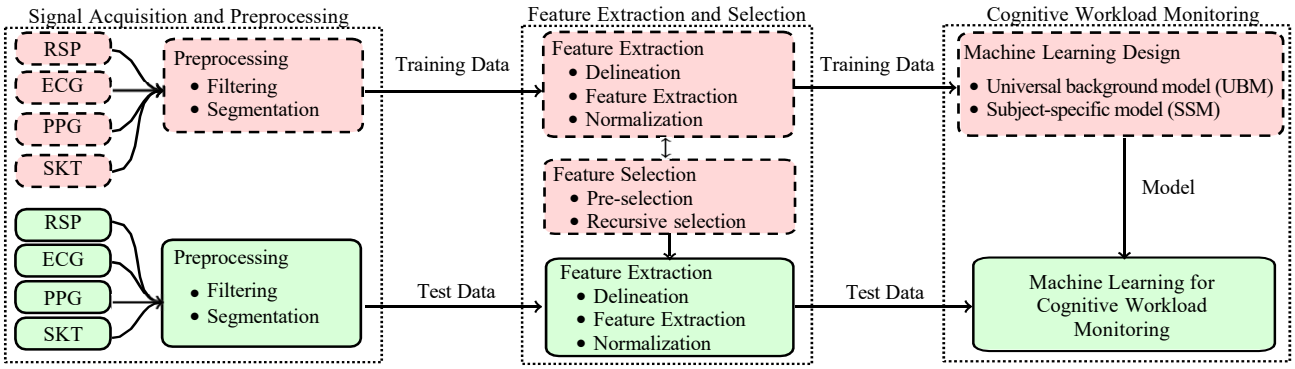

Fig. 1. Process overview for the design of a CWM method. Blocks with dashed lines represent the applied design/optimization methods and blocs with solid lines represent the final system.

III. CWM System

The general design of an ML algorithm suitable to develop a wearable embedded system for online CWM is shown in Fig. 1, namely, blocks with solid lines. Instead, the blocks with dashed lines represent the different statistical pattern recognition methods applied to experimental data for designing such a system. All our analyses are done offline, but the final system is tested, emulating online processing (i.e., using causal spectral filters and computing the features using past information).

The system is divided into three main steps shown in Fig. 1 with the dotted lines, i.e., Signal Acquisition and Preprocessing, Feature Extraction and Selection, and CWM. In the first step of the CWM system, sliding window is applied for signal segmentation, which defines the time resolution of the workload monitoring system. The preprocessing consists of removing artifacts from the signals. In this work, we collected experimental data for both design and evaluation of the proposed CWM method.

Next, the features extraction and selection step includes generating a feature vector that best represents the physiological response induced by different workloads. For an exhaustive investigation, we chose an exploratory approach in which we extract a large number of different features in both time and frequency domains. Then, since physiological signals exhibit high intra- and inter-subject variability due to age, gender, time of day and other factors [30], we investigate different normalization methods. Subsequently, we apply different features selection methods to define the best subset of features to be used in the final system.

Finally, the CWM step includes the prediction of a discrete CWL level. For the design of the CWM method, we consider the most common machine-learning techniques based on pattern recognition algorithms suitable for implementation in embedded systems. Moreover, we consider a personalized weighted-learning approach to assess the person-dependent variance in the physiological response of an induced workload. Performance of our method is then evaluated based on NASA Task Load Index (NASA-TLX), a subjective and multidimensional assessment tool that rates perceived workload [7].

IV. Signal Acquisition and Preprocessing

For a thorough exploration of the physiological changes induced by cognitive workload, we measure RSP, ECG, PPG, SKT, EDA, and EEG, which are signals that are typically used in the literature [34], [35]. The effect of cognitive workload on EEG was analyzed and presented in a different work [36]. Here, we focus on the remaining signals, which sensors can be integrates into a wearable system, such as FlyJacket [17].

Their main physiological manifestations related to CWL are reported in Table II and described in Sec. IV-A.

A. The physiological process behind CWL

While performing a very demanding task, the need for more oxygen is driven by the autonomic nervous system (ANS) activation. The latter involves both a sympathetic nervous system (SNS) activation and parasympathetic nervous system (PSNS) counterbalance. This increased oxygen demand triggers faster and deeper respiration [37]. Therefore, RSP should be measured to track CWL changes [20].

The ANS activation also triggers a cardiac response, which is also affected by the Hypotalai-Adrena (HPA) axis. This response is associated with variabilities in heart rate, defined as heart rate variability (HRV) obtained by monitoring the ECG signal. Consequently, the above relationship can explain the heart's ability to respond to multiple physiological and environmental stimuli [8]. The neurohypophysis activation, the HPA axis, and the ANS lead to blood volume changes, peripheral blood vessels resistance, and cardiac response derived from the pulse wave. Features from the PPG are used to detect those physiological changes induced by cognitive tasks [24], [37].

Moreover, it has been proved that cognitive tasks cause peripheral vasoconstriction [24], [37], regulated by the vasoregulatory system and driven by both neurohypophysis and SNS. Thus, SKT is required to detect the variations in peripheral temperature that are associated with peripheral vasoconstriction.

Finally, EDA is one of the most commonly used measures in studies involving emotional arousal. According to [38], EDA is traditionally measured at the fingers or palms, while foot and shoulders seems to be valid alternatives for ambulatory measurement. However, we cannot confirm their findings, as our EDA measurements from the shoulder did not show any significant response. Therefore, EDA measurements were not considered in this work.

B. Signal preprocessing

The first preprocessing step consists of removing the artifacts from the signals with causal filters [16]. We apply a baseline wander with cutoff frequency at 0.3 Hz to both ECG and PPG signals. Next, we also apply a 32nd-order bandpass

TABLE II
PHYSIOLOGICAL MANIFESTATIONS RELATED TO INDUCED CWL.

Physiological measures	Measurable physiological manifestation to workload response	Sensor body position
Peripheral skin temperature	Neurohypophysis and Sympathetic Nervous System (SNS) activation	Finger
Respiration	SNS activation and Parasympathetic Nervous System (PSNS) counterbalance	Thorax
Electrocardiogram	Both Hypotalai-Adrena (HPA) axis and SNS activation, and PSNS counterbalance	Thorax
Photoplethysmography	Neurohypophysis, HPA axis, and SNS activation, and PSNS counterbalance	Ear

FIR filter with linear phase and Hamming window with cut-off frequencies at 0.3 and 30 Hz for ECG and at 0.1 and 5 Hz for PPG [37]. In the case of the RSP signal, we employ a 4th-order Butterworth IIR bandpass filter with cutoff frequencies at 0.03 and 0.9 Hz. Nevertheless, because of the slow response time of the SKT thermistor (1.1 sec.), which avoid the high frequency noise, no filter is applied to the acquired SKT signal.

Finally, we apply a time-series segmentation of all the acquired physiological signals, which are thus divided into a sequence of samples in windows of 60 seconds.

V. FEATURES EXTRACTION AND SELECTION

Following our methodology described in Section III, we perform an offline investigation to select the features to be considered in the final system. That is, we first extract a broad features set from the segmented signals for an exhaustive assessment of the person's physiological response to CWL. Then, we select the best features set rich in discriminatory information concerning the physiological states induced by different CWL levels. normalized and given as input to the developed CWM algorithm.

A. Feature extraction

For the design of the CWM system, our feature extraction process includes three main steps. First, we delineate the segmented signal to detect points of interest (e.g., signal onset, peak, offset, etc.). Second, we extract physiological markers, a combination of different delineated points and provide information about the person's physiological state (e.g., heart rate). Finally, we compute features in both time and frequency domains. For the time domain, we use standard statistical features (i.e., mean, median, mode, standard deviation, variance, root mean square, and power), extracted either from the physiological markers or from the segmented signals directly. However, in the frequency domain, the features are computed specific to the characteristic of the physiology of each signal, which are listed and detailed next.

Following an extensive literature review and by applying our experience from previous projects [8], we increased the number of analytical methods applied to a single physiological signal segment to extract 384 features: 127 from RSP, 38 from ECG, 190 from PPG, 2 from SKT, and 27 from RSA. However, applying our feature selection method, the final system uses only 25 features, 10 from RSP, 2 from ECG, 10 from PPG, 2 from SKT, and 1 from RSA. These 25 features are listed in Table IV. From EDA, we aimed to compute the mean skin conductance level and the number of skin conductance responses per minute as in [38]. Though we used dedicated electrodes (recommended by Biopac), our EDA signal was rudely flat across participants suggesting a poor SNS activation on the shoulders for our study case. Thus, the signal was discarded. More details about the delineation and feature extraction for each considered signal are provided next.

Fig. 2 shows a schematic representation of the signal processing and feature extraction process.

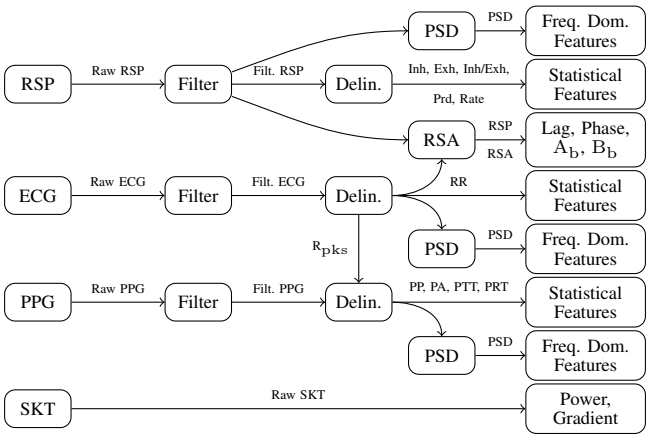

Fig. 2. Schematic representation of the signal processing and feature extraction processes.

1) Respiratory activity (RSP): To extract the features from the RSP signal, we first delineate the signal based on the differences between adjacent samples of the filtered signal defined as:

$$\Delta x[k] = x[k] - x[k-1] \tag{1}$$

Then, by comparing both current and previous values, we detect from the sign of Δx the falling and rising edge, which coincide with inhalation (RSP-peaks) and exhalation (RSP-valleys) end, respectively. Then, all peaks and valleys pairs having a difference smaller than 20% of the mean RSP amplitude are removed [31].

Next, from the delineated RSP, we extract the following physiological markers: inhalation (Inh) and exhalation (Exh) time, the Inh/Enh ratio, Inh and Exh amplitudes, respiratory period ($\mathrm{RSP}_{\mathrm{Prd}}$), and respiratory rate ($\mathrm{RSP}_{\mathrm{Rate}}$). Besides, we compute their numerical differences using Eq. 1. Finally, we calculate the segmented RSP signal's statistical features, its difference (Eq. 1), and all the aforementioned RSP physiological markers. In the frequency domain, we compute the power of the segmented signal in four different bands of equal bandwidth (i.e., 0-0.25, 0.25-0.5, 0.5-0.75, and 0.75-1 Hz), as reported in [28]. We also consider the normalized band powers, obtained by dividing each of the above band powers by the total power in the 0-1 Hz band.

2) Electrocardiogram (ECG): We compute the so-called normal-to-normal (NN) intervals from the filtered ECG signal, the intervals between normal QRS complexes detected with the delineation method described in [39]. Then, we compute features in the time domain describing the Heart Rate Variability (HRV) [40], which are statistical features of the successive NN-intervals and of the interval differences of successive NN-intervals. We also computed the number of interval differences of successive NN-intervals greater than 50 ms (NN50) and the proportion derived by dividing NN50 by the total number of NN-intervals (pNN50) within the processing window.

Additionally, we obtain several geometrical features from the Poincaré (or Lorenz) plot indicating vagal and sympathetic

functions. In particular, we extract the length of the transverse axis (T), vertical to the line $\mathrm{NN}_k = \mathrm{NN}_{k+1}$; the length of the longitudinal axis (L), parallel with the line $\mathrm{NN}_k = \mathrm{NN}_{k+1}$; the Cardiac Sympathetic Index (CSI), defined as L/T; the modified CSI (L^2/T); and the Cardiac Vagal Index (CVI) as $log_{10}(LT)$ [40].

Moreover, we extract HRV features from the frequency-domain, as proposed in [40]. That is, the power in two frequency bands, namely, low-frequency (LF: between 0.04 and 0.15 Hz) and high-frequency (HF: between 0.15 and 0.4 Hz). LF and HF powers are obtained from estimating of the Lomb-Scargle Power Spectral Density (PSD) of the NN intervals [41]. The power values are divided by the total power minus the very-low-frequency (VLF) component (frequency $\leq$ 0.04 Hz). Also, we compute the power sum LF + 1/HF and the ratio LF/HF.

Furthermore, we extract novel features from the HF band. The first one, called $\mathrm{RR}_{\mathrm{HF\ gauss}}$, is the mean frequency of a Gaussian distribution used to fit the Lomb-Scargle PSD estimated in the HF band. This feature describes the shifting in frequency of the PSD in the HF band, where the shift is mainly caused by the RSP activity [42]. The second one is called $\mathrm{RR}_{\mathrm{HF\ pond}}$ and is defined as:

$$\mathrm{RR}_{\mathrm{HF\ pond}} = \frac{\sum_{f\in \mathrm{HF}} f\ \mathrm{PSD}\{RR[k]\}(f)}{\sum_{f\in \mathrm{HF}} PSD\{RR[k]\}(f)} \tag{2}$$

Finally, we also compute the power of the HF divided in 5 sub-bands of equal length ($\mathrm{RR}_{\mathrm{HF\ sband\ Xn}}$), where the subscript index $X = \{1, \cdots, 5\}$.

3) Photoplethysmogram (PPG): According to [37], we delineate the PPG signal and extract the following physiological markers: the Pulse Period (PP), the time interval between two consecutive pulse peaks; the Pulse Amplitude (PA), the difference between the pulse peak and the pulse onset; the Pulse Transit Time ($\mathrm{PTT}_{\mathrm{M}}$), the time interval between the R-Peak in the ECG signal and the instant when the PPG pulse reaches half of its onset-to-peak amplitude; the Pulse Rise Time (PRT), the time interval between the pulse onset and the pulse peak; and the Pulse Rise Speed (PRS), the ratio between amplitude difference and time interval computed from the pulse wave points located at 75% and 25% of the onset-to-peak amplitude, respectively.

To have accurate estimations of PTT and PRT, in the literature [16], the use of both ECG and PPG signals has been proposed. Using both enables trade-offs between accuracy and complexity of the sensing wearable system.

From each of the aforementioned PPG physiological markers, we extract features in the time and frequency domains, following the HRV methodology applied to NN-intervals.

4) Peripheral Skin Temperature (SKT): From the SKT signal, we directly extract the $\mathrm{SKT}_{\mathrm{Gradient}}$ and $\mathrm{SKT}_{\mathrm{Power}}$ of the signal. The SKT Gradient is computed as the mean of the difference between the portion of samples recorded during the first second of the window, acquired at a sampling frequency f_s, and the samples from the final one second of the window. Then, the SKT Power is the signal average power of computed over the entire window of samples.

5) Respiratory Sinus Arrhythmia (RSA): Respiratory sinus arrhythmia (RSA) is the natural variation in the heart rate associated with the respiratory cycle. and measured from the ECG signal. RSA has been used as a noninvasive measure of cardiac vagal tone, as a marker of PSNS tone [43] and thus, it can be used as a marker of the disruption of homeostasis induced by a highly demanding task. Since RSA and cardiac vagal tone can dissociate under certain circumstances [44], we consider the hypotheses that these differences could come from external factors, such as, a need to compensate for CWL changes.

RSA is estimated from the non-uniform time series of successive NN-intervals, which we interpolate using a linear function and resample at 2 kHz. Then, we filter the resulting uniform time series of successive NN-intervals with a 4th order band-pass Butterworth filter with cutting frequency at 0.15 and 0.4 Hz yielding a RSA.

From the computed RSA we extract features that aim to evaluate the agreement with the measured RSP signal, but first, both signals (RSP and RSA) are normalized to zero mean and unit variance. The first feature is the time delay of the RSA with respect to the RSP ($\mathrm{RSA}_{\mathrm{Lag}}$), estimated by computing the cross-correlation of RSA and RSP. We also compute the phase shift between the two signals, given by Eq. 3.

$$\mathrm{RSA}_{\mathrm{Phase}} = \cos^{-1}\left(\frac{\mathrm{RSP}\cdot\mathrm{RSA}}{\|\mathrm{RSP}\|\cdot\|\mathrm{RSA}\|}\right) \tag{3}$$

Subsequently, we extract features based on the Tukey mean-difference plot, also called the Bland-Altman plot [45], to compare both RSA and RSP measurements. To this end, we compute the statistical features of the difference between the two signals and the mean of the two:

$$R_0 = \mathrm{RSP} - \mathrm{RSA} \tag{4}$$

$$A_0 = (\mathrm{RSP} + \mathrm{RSA})/2 \tag{5}$$

We also consider the statistical features of different log transformations of the measurements, as follows:

$$R_b = log_b(\mathrm{RSP}) - log_b(\mathrm{RSA}) \tag{6}$$

$$A_b = (log_b(\mathrm{RSP}) + log_b(\mathrm{RSA}))/2, \forall b = \{n, 2, 10\} \tag{7}$$

where b denote the logarithm base (i.e., n, 2, and 10).

B. Features Normalization

Since the relative range of each feature varies widely, a normalization is applied so that each one contributes approximately equally to the classification problem. Hence, we apply a min-max normalization scaling the features within a 0-1 range. The general formula is given as:

$$\mathbf{x}' = \frac{\mathbf{x} - \min(\mathbf{x}^\dagger)}{\max(\mathbf{x}^\dagger) - \min(\mathbf{x}^\dagger)} \tag{8}$$

where $\mathbf{x}$ is an original value, $\mathbf{x}'$ is the normalized one, and $\mathbf{x}^\dagger$ represents the original value of the training set.

Moreover, to address the problem related to both inter-subject and inter-day variability [8], [30], we found from the computational vision community, a task-specific normalization method [46], which inspired us to consider the following three types of normalization. First, the total normalization (TN) is based on the full training set. Second, the subject dependent normalization (SN) consists on normalizing based on each training subset relative to a specific subject. Finally, the day and subject-dependent normalization (DSN) affects each portion of the training set relative to a specific day and subject. Thus, the training and the test sets are scaled accordingly, using the parameters obtained only from the training set.

Finally, we select the best normalization strategy that better emphasizes the discriminant power of the features and their ability to classify the problem. In other words, we select the method that gives the highest Fisher Discriminant Ratio (FDR) [47] of the normalized feature sets, obtained by applying one

of the three different normalization methods (i.e., TN, SN, or DSN). Then, we evaluate the classification performance of an SVM that uses for each normalized set an equal number of normalized features. The results are reported in Sec. VIII-B.

C. Features Selection

Given the large features number considered for the exhaustive characterization of CWL, we divide the feature selection process into two main steps. First, as a pre-reduction to suppress the features that do not give any discriminatory information, we apply filter methods, particularly effective in computation time and robust to overfitting. Then, to select the most important features considering their possible interactions, we apply embedded methods that simultaneously perform feature selection and classification. Both feature selection steps are performed once with data from the training set.

The pre-reduction of the feature space involves three methods. First, a two-sample Student's t-test selects statistically discriminant features. Second, the discriminant features are ranked based on their FDR, which gives a score based on their ability to discriminate the problem. Lastly, we remove the features that give any redundant information, the less discriminant features that are strongly correlated with others (i.e., a Pearson's correlation coefficient above 0.95) [48].

For the final feature selection, we apply Recursive Features Elimination (RFE) [49], an embedded method that uses an external estimator to assign weights to features. These weights are then used to prune the least important features from the current set. This procedure recursively prunes the selected features until all feature weights are different from 0. In this work, we apply RFE based on different classifiers (i.e., LR, LDA, SVM, RF, and XGB), which we name RFE-LR, RFE-LDA, RFE-SVM, RFE-RF, and RFE-XGB, respectively.

VI. Cognitive Workload Monitoring

For the cognitive workload monitoring, we explore the use of different machine-learning algorithms. In particular, we investigate the use of linear models, namely LR, LDA, SVM, and Gaussian Naive Bayes (GNB) for a feasibility check. Then, we investigate the use of non-linear models, such as k-Nearest Neighbour (k-NN), Quadratic Discriminant Analysis (QDA), SVM with a Radial Basis Function (RBF) kernel, DT, RF, and XGB, to reduce the bias. The accuracy of each model in detecting high levels of CWL is evaluated based on a 5-fold cross-validation (CV) over the training set.

Moreover, we consider a personalized weighted-learning approach to deal with the person-dependent variance. To this aim, we compare the performance of the Universal Background Model (UBM) and the Subject-Specific Model (SSM) [50].

A. Model for Cognitive Workload Monitoring

To estimate CWL, we chose a linear SVM that has the following prediction model [51]:

$$y(x) = \mathbf{w}^T\mathbf{x} + b \tag{9}$$

where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is the weight vector, and b is the offset. The corresponding optimal hyperplane separating the two classes is defined by the relation:

$$y(x) = \mathbf{w}^T\mathbf{x} + b = 0 \tag{10}$$

Thus, an input vector $\mathbf{x}$ is then assigned to class 1 if $y(x) \geq 0$ and to class -1 otherwise. Although we use the same prediction model for UBM and SSM, the difference lies in the objective function. All the details are given in Sec. VI-B and VI-C.

The parameters of both UBM and SSM are chosen based on a 5-fold CV on the training set. We use a stratified split for this validation that preserves the same percentage for each target class as in the complete training set and preserves the same percentage of data relative to the subject of interest. Then, the generalization of both models is tested on an unseen test set.

The performance of the models is evaluated based on: accuracy, the proportion of both true positives and true negatives results among the total number of cases; precision, or confidence, the proportion of predicted positive cases that are correctly real positives; recall, or sensitivity, the proportion of real positive cases that are correctly predicted positive; Receiver Operating Characteristic (ROC); and in particular, based on the F1-score, the weighted average of the precision and recall.

B. Training of the Universal Background Model

The considered UBM is based on SVM with soft margins [51], which relax the condition for the optimal hyperplane (Eq. 10) and allow possible overlaps of the class-conditional distributions. As for a normal soft-margin SVM, the objective function of the UBM is defined as follows:

$$\begin{aligned} \underset{w,b,\xi_i}{\arg\min} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i\in D}\xi_i, \\ \text{subject to} \quad & t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0; \quad (i \in D) \end{aligned} \tag{11}$$

where the regularization term C and the non-negative variables ξ_i relax the constraints of an otherwise hard-margin SVM. The data x in the training dataset D comprises N input vectors $x_1, \cdots, x_N$, with corresponding target values $t_1, \cdots, t_N$, and where $t_i \in \{-1, 1\}$. The parameter C is analogous to the inverse of a regularization coefficient because it controls the trade-off between minimizing training errors and controlling model complexity. A regularization term $C = 0.1$ is chosen from a $\log_{10}$ scale ranges from 0.001 to 1000 based on a stratified 5-fold CV on the training set.

C. Training of the Subject-Specific Model

As well as for the UBM, the considered SSM is based on a soft-margin SVM. However, to adapt the model to a specific subject, we modify the objective function of the original soft margin SVM (Eq. 11) including two different soft-margins. The first soft-margin (C_s) changes the importance degree given to false estimations of samples coming from a particular subset of data, which can be a particular subject (S). Thus, the term weighed by C_s allows a minimization of the errors (ξ) for all the x in the training set related to a specific subject ($x \in S$). Instead, the second soft-margin (C) affects the rest of the dataset minimizing the errors ξ for all the x in the training set that are related to other subjects ($x \notin S$).

Therefore, the SSM final objective function is defined as:

$$\begin{aligned} \underset{w,b,\xi_i}{\arg\min} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i\notin S}\xi_i + C_s\sum_{i\in S}\xi_i \\ \text{subject to} \quad & t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0; \quad (i \in D) \\ & C_s > C \end{aligned} \tag{12}$$

With this model, we state a preference for margins that classify the training data correctly, but we soften the constraints to allow for non-separable data with different penalties. To promote the minimization of the total sum of the penalties

$\xi_i \ \forall i \in S$, despite the minimization of the total sum of the penalties $\xi_i \ \forall i \notin S$, we chose C_s to be greater than C. As usual, the regularization terms have to be large enough to avoid under-fitting, but not too much to avoid over-fitting as well. Based on a stratified 5-fold CV on the training set, both regularization terms $C = 0.001$ and $C_s = 0.1$ are chosen from a $\log_{10}$ scale in ranges of 0.001–0.1 and 0.1–100, respectively. Although both regularization terms seem to be bounded by the considered range, we keep the lower bounds to avoid possible under-fitting problems.

VII. Experimental setup

Collecting data in a real SAR mission is complex because of the random frequency of events and the many variables still undefined. Therefore, for collecting clean data, building a CWM model, and validating our approach, we used the simulator for search and rescue mission with drones reported in [8]. With the help of a certified instructor of the Swiss firefighters, we designed the following two study protocols, where both are based on a repeated-measures design using counterbalancing. The first study was conducted to characterize CWL levels through physiological signals using a gamepad as controller, to build a model for real-time monitoring, and to evaluate the contribution of the subject-specific approach.

The second protocol was designed to evaluate the system's quality using a new advanced controller, the FlyJacket [17]. In contrast with the gamepad controller, where the movements were limited to the thumbs, the FlyJacket implies both arms and torso movements. Therefore, when comparing tasks involving different types of movements, there is a risk of yielding a performance overestimation. Thus, to avoid as much as possible any possible miss-classification caused by movement artifacts, we trained the machine-learning algorithm with the data from Study 1 (Trial 1) with the gamepad and did the SSM final tuning with data from Study 2 with FlyJacket. Finally, our models were tested also on unseen data of Study 2 with FlyJacket. The details of both studies are in the following sections.

The signal processing, features extraction, machine-learning design, and classification were done using Matlab R2016a [52]. The RSP, ECG, PPG, SKT, and EDA were recorded with the Biopac MP160 system at 2 kHz of sampling frequency. We also recorded EEG, but because of the difficult integration of such a sensor into a jacket, it is not used in this work. Instead, it is analysed in [36], as previously mentioned. Finally, through an analog input of the Biopac system, a trigger signal provided by the simulator advises the task execution.

A. Search and rescue drone simulator

As presented in [8] and [36], the simulator presents a simplified SAR scenario, where the drone pilot has to deal with two different activities, flying and mapping. The flying activity consists in flying a drone following a randomly generated trajectory depicted by spherical waypoints. Instead, the mapping activity consists of mapping a disaster area situation, represented by cubes of 4 different colors randomly distributed over the flying trajectory. The colors were chosen according to the regulation of the Swiss Firefighters [53].

We modulate both flying and mapping activities to induce different levels of CWL as in [8], [36] i.e., medium/high workload level with Flying (F) and Mapping 3 objects (3M), and high level of CWL with Flying and Mapping 3 objects (F3M). Also, a flying sequence controlled by an auto-pilot is used as Baseline (B) to have participants in a same framework for the entire experiment. B task has the lowest expected workload level of this study.

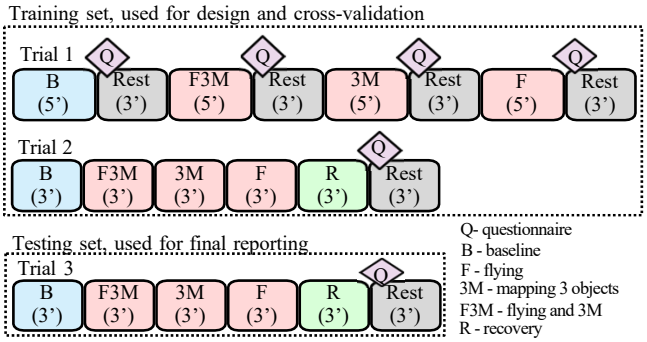

Fig. 3. Protocol of the experiment with the gamepad.

B. Study protocol 1: Use of a gamepad

During this study, participants sat in front of a screen and controlled the simulator with a gamepad from Logitech. To collect clean data, participants were asked not to talk and to avoid any kind of unnecessary movements during the tasks. For proving the feasibility of detecting cognitive workload with constrained sensor placement, clean data were needed. Hence, we asked the participants not to talk and avoid unnecessary movements while performing the tasks. However, we cannot completely avoid the presence of some artifacts. Therefore, in this context, different methods can be applied to make sure the input data can be used for our proposed algorithm. In particular, different approaches in wearables have been shown to be effective for noise removal (e.g., for speech [54], [55] and movement [56], [57] artifacts), which are needed in real-life scenario. The study started with a setup phase (explanation about the experiment, request of the participant consent, and sensor placement), followed by a warm-up phase up to 10 minutes to get familiar with the simulator [58].

The study protocol is shown in Fig 3. Participants performed the first trial, starting with a five-minute baseline, and followed by a sequence including F3M, 3M, and F, executed in a randomized order. A resting period of 3 minutes was enforced after each task. This period also allowed participants to fill a questionnaire (Q), based on the NASA-TLX procedure.

Finally, the participants performed two additional trials, namely Trial 2 and Trial 3. Each trial started with a baseline and continued with a randomized sequence of F3M, 3M, and F, and ended with a recovery (R) phase followed by a resting period, in which the NASA-TLX was filled again. Each task presented in Trial 2 and 3 lasted three minutes.

As shown in Fig. 3, we used all data acquired during both Trial 1 and Trial 2 for both training and CV, and all data collected during Trial 3 as the final unseen test set. We are conscious that this split does not truly respect independent temporality of data because all data sets (i.e., training, CV, and unseen test sets) are taken from the same day and not from a day that is not used for testing (as it should be in a real application). Therefore, this choice implies a daily training phase, which can be seen as a daily calibration of the system. However, as we expect an inter-day variability of the physiological responses [8], [30], we assume that a daily calibration of the system will be required. This calibration process consists of tuning the model for the correct baseline level by using a couple of minutes of data collected under both low and high workloads. A further investigation over different

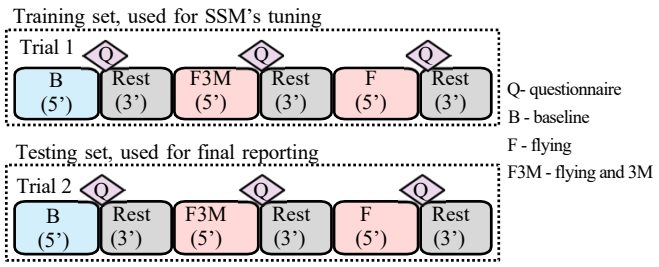

Fig. 4. Protocol of the experiment with FlyJacket.

days could potentially avoid the need for such a calibration, but this analysis is left for a future study.

C. Study protocol 2: Use of FlyJacket

In this study, the drone simulator is controlled with the FlyJacket and two Oculus Touch controllers to map the disaster situation. The study also started with a setup and warm-up phase. Then, participants performed two trials, as shown in Fig. 4, which started with a five-minute baseline followed by a F3M and F sequence executed in a randomized order. Again, three-minute resting period was enforced after each task, where the participants filled the questionnaire.

This second study is a reduced version of the first one since it aims to prove the feasibility of detecting low and high CWL levels with the proposed method. Hence, we designed this study protocol with only two trials, with three tasks of five minutes each, and recording F for a different study [15].

D. Research participants

Study 1 with the gamepad was done by 24 participants (6 females and 18 males) aged between 21 and 39 years old (27.7 ± 4.8), who performed the study protocol twice in two sessions on different days. Study 2 with the FlyJacket was done by 10 additional participants (3 females and 7 males) aged between 22 and 30 years old (26.8 ± 2.3), on a single day session. All participants provided informed consent to participate in both studies. The inclusion criteria were being healthy, free of any cardiac abnormalities, and were receiving no medical treatment. The Cantonal Ethics Commissions approved this study for Human Research Vaud and Geneva (PB2017-00295).

VIII. Experimental Results

Given the recorded data set from Study 1, we select the best combination of normalization, feature selection, and classification methods suitable for CWM. The methods are obtained based on the cross-validations workflow including 747 observations. Finally, we show the performance of the proposed methods on two unseen test sets, including 260 and 57 observations from Study 1 and 2, respectively.

A. Self-perception of induced cognitive workload

The reported overall workload on each task perceived by the 34 participants based on the NASA-TLX is shown in Fig. 5. A one-way ANOVA conducted on the influence of the tasks confirms that participants have perceived different levels of workload. Furthermore, a multiple pairwise comparison analysis using the Student's t-test with up to 164 samples revealed statistically significant mean differences, except for 3M vs F (p-value < 0.001). The comparisons with the 3M task

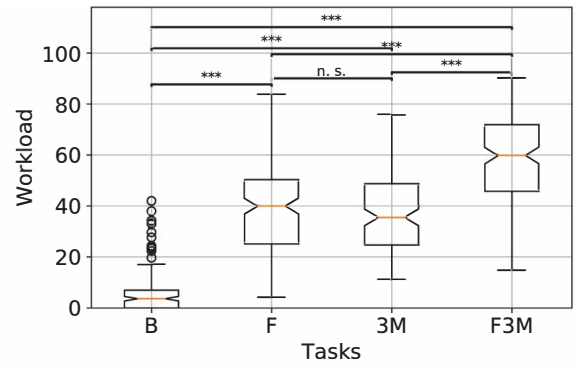

Fig. 5. Cognitive workload perceived by participants.

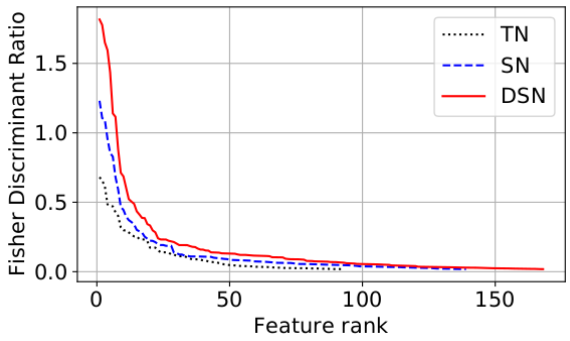

Fig. 6. Normalization methods impact on FDR.

were limited to 144 samples, as Study 2 with the FlyJacket setup does not include the 3M task.

However, as shown in Fig. 5, the perceived CWL level has a large variance. A two-way ANOVA reveals that such a large variance comes from a significant ($p < 0.001$) effect of task, day, and subject on the level of CWL, F(3,414) = 1637.19, F(1,414) = 28.70, F(33,414) = 48.93, respectively. Therefore, the NASA-TLX results confirm the need for both a day- and a subject-specific approach.

Although there is a significant difference in the perceived workload between most tasks, Fig. 5 shows that the distribution of both F and 3M presented a considerable overlap with F3M. Instead, the difference between tasks B and F3M is clear. Thus, as our main goal is to detect low and high levels of CWL, we focus on the extreme cases induced by tasks B and F3M, respectively. F and 3M conditions were analysed in a different work [15], which targets a three-class CWM.

B. Features discriminant power emphasized by normalization

To reduce the variance introduced by the different participants and performing the experiment on different days, we investigated different normalization approaches (i.e., TN, SN, and DSN) as described in Section V-B. We firstly evaluated the effect of each normalization approach on the features discriminant power based on their FDR. Results are shown in Fig. 6, where DSN better emphasises the discriminant power of the features. Compared with TN, the FDR of the most important feature is emphasized by a factor of 80.9% or 166.9%, over SN or DSN, respectively.

Secondly, following our methodology (see Sec. III), we compare how each normalization approach contributes to the classification problem using a linear SVM model. We noticed that the normalization affects the feature selection process,

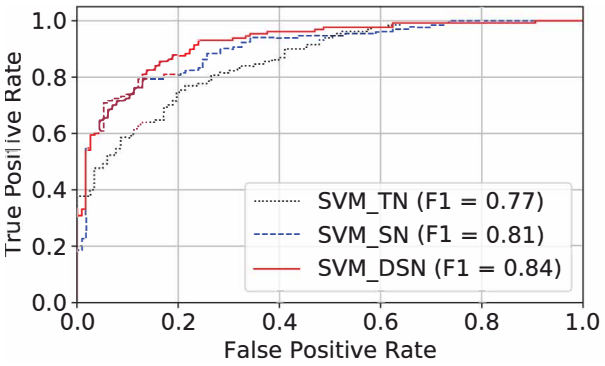

Fig. 7. Normalization methods impact on CWM.

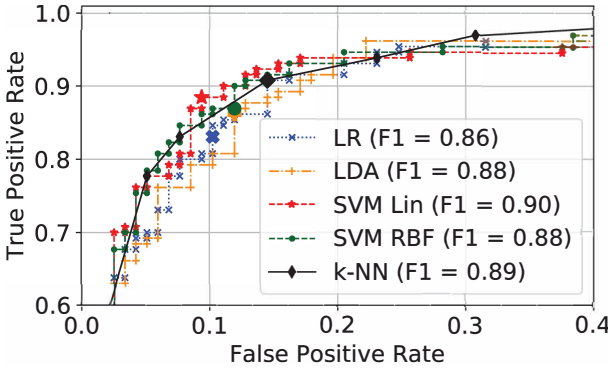

Fig. 8. Best classifiers comparison on CV. Bigger markers denote the performance of the different models based on their corresponding cross-validated threshold or offset $\boldsymbol{b}$.

which selects 14 features after TN or SN, or 25 features after DSN. Therefore, to avoid biased results caused by the use of a different number of features, we used for this comparison the first 14 most discriminant features selected by RFE-SVM after TN, SN, or DSN normalization. Fig. 7 shows the ROC and the F1-score of the SVM combined with the different normalization methods, where it can be seen that once again DSN outperforms both TN and SN.

Our results show that feature normalization plays an important role during both features selection and classification. DSN normalization gives better results (a bigger F1-score) compared to SN and TN. Similar trends are obtained by applying RFE with other classifiers, such as LR or LDA. Therefore, we select DNS as normalization method.

C. Physiological featuring of cognitive workload

By applying the filter methods presented in Section V-C, we eliminated 282 non-informative features from the normalized (based on DSN) 384 features initially considered for an exhaustive CWL characterization. In particular, we reduced the feature space dimension from 384 down to 168 features with the two-sample Student's t-test and down to 102 features by checking their linear correlation.

Although the above pre-selection step drastically reduced the feature space, using that amount of features requires models with high capacity. It may lead to overfitting if trained with a limited dataset like ours. Therefore, to obtain a reasonable feature set that can be used for CWM, a further dimension reduction based on embedded methods was applied, as presented in Section V-C.

The features space was reduced from 102 to 5, 10, 12 and 25 by applying RFE-XGB, RFE-LR, RFE-LDA, RFE-SVM, respectively. RFE found a consistent set of features based on LR, LDA, and SVM, see Table III. For the case of RFE-XGB, we used a low-complex model to avoid overfitting and inconsistent results. In particular, we limited the model to 10 estimators and three maximum depth of each decision tree. Such a low-complex RFE-XGB showed a drastic lower selection compared to other methods.

Without banning the ensemble methods from building complex models, RFE does not converge to the same result if executed several times. In contrast, by limiting the model complexity, RFE provides a reproducible result. However, this trick does not help the RFE-RF method that does not converge to a consistent solution. This model always selects a different set of features, even if the model complexity is reduced (i.e., number of estimators and maximum tree depth). Hence, such complex models are not suitable for small datasets.

The feature set obtained after applying both filter and embedded methods are shown in Table IV. Although selected features vary between 5 and 25 depending on the applied embedded methods, a common subset of features is identified. We observed that the features obtained by RFE-LR, RFE-LDA and RFE-XGB are almost all included in the feature set obtained by RFE-SVM. In particular, $\mathrm{RSP}_{\mathrm{Rate\ Median}}$ and $\mathrm{SKT}_{\mathrm{Power}}$ are selected by all the four methods, followed by $\mathrm{RSP}_{\mathrm{Prd\ Median}}$, $\mathrm{SKT}_{\mathrm{Gradient}}$, $\mathrm{RSA}_{\mathrm{R2\ Std}}$, $\mathrm{PRT}_{\mathrm{Median}}$, $\mathrm{RSP}_{\mathrm{Rate\ Diff\ RMS}}$ and $\mathrm{PP}_{\mathrm{Median}}$, selected by three methods out of four. Based on this result, the above eight features seem to be the most important ones in terms of CWL characterization in the context of this experiment.

Additionally, we investigated the effect of using the different feature sets obtained with the considered RFE methods on different classification methods. Results are presented in Table III, where we report both the training and the CV accuracy. A significant difference between training and CV accuracy indicates a sign of overfitting (e.g., QDA with 102 features). Moreover, we report the best CV F1-score for each applied RFE method. While there seem to be no significant differences across methods, the highest best F1-score and the best CV accuracy are reached when linear SVM is applied on both RFE and classification. Therefore, RFE-SVM is the employed feature selection method hereafter.

D. Classifiers for cognitive workload monitoring

A ROC curve is used to further evaluate the performance of the considered classifiers in CV, reported in Fig. 8 . In particular, for greater clarity of the illustration, we only report the best classifiers results ($\mathrm{AUC} \geqslant 0.94$), namely LR, LDA, k-NN, linear SVM, and SVM with RBF kernel. Our results show that, with the amount of data we have, the use of non-linear models does not increase the detection accuracy. Instead, non-linear models tend to introduce a larger variance between training and CV- accuracy. Linear SVM shows a higher F1-score and better ROC curve, in particular by comparing the bigger markers representing the performance of the models based on their corresponding cross-validated threshold or offset b. Therefore, a linear SVM was selected for our further investigation.

Although selecting the SVM reaches the highest classification accuracy, it may not be the optimal solution for embedded implementations. Other solutions considering fewer features may be preferred for implementations in low-power embedded systems, where power consumption may play an important role. However, our results indicate certain flexibility in selecting the number of features to be used, since the best F1-score is quite similar for all the applied feature selection embedded methods.

TABLE III
FEATURE SELECTION PERFORMANCE COMBINING RECURSIVE FEATURE ELIMINATION (RFE) AND CLASSIFICATION METHODS.

Embedded Feature Selection		Training and cross-validation accuracy of different classifiers, left and right values, respectively																				F1-score
Method	Features	LR		LDA		QDA		SVM_{Lin}		SVM_{RBF}		GNB		k-NN		DT		RF		XGB		Best
Pre-selection	102	93	85	94	84	100	82	88	87	92	87*	84	85	91	85	90	85	90	87	94	87	89
RFE-XGB	5	85	85	85	86	87	86	85	87*	87	88*	85	88*	90	87*	89	84	90	85	90	85	88
RFE-LR	10	90	86	91	85	90	87*	89	86	91	86	86	86	91	85	90	80	90	85	92	84	88
RFE-LDA	12	90	87	91	87	92	87	87	86	92	88	86	85	92	87*	88	76	90	85	91	85	89
RFE-SVM	25	91	86	92	87	95	88	89	88*	93	87	88	88	92	87	90	84	92	85	93	85	90

* highlights the classifier having the best F1-score on cross-validation for the particular feature selection method. Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), with linear and Radial Basis Function (RBF) kernels, Gaussian Naive Bayes (GNB), Nearest Neighbour (k-NN), Decision Tree (DT), Random forest (RF), and Extreme Gradient Boosting (XGB).

TABLE IV
MOST IMPORTANT FEATURES USED TO DETECT LOW AND HIGH LEVELS OF CWL, B AND F3M TASKS, RESPECTIVELY.

Physiological Features	Task B $\mu \pm \sigma$	Task F3M $\mu \pm \sigma$	p-Val$< 10^{-x}$ x
$RSP_{\text{Rate Mean}}$ [1,4]	0.28 ± 0.22	0.71 ± 0.23	107
$RSP_{\text{Rate Median}}$ [1,2,3,4]	0.28 ± 0.23	0.71 ± 0.23	106
$RSP_{\text{Prd Mean}}$ [1,2,3]	0.61 ± 0.25	0.22 ± 0.21	90
$Inh_{\text{Time Median}}$ [1]	0.53 ± 0.31	0.20 ± 0.23	52
$Exh_{\text{Time Median}}$ [1]	0.63 ± 0.29	0.32 ± 0.28	45
$Inh_{\text{Time Mean}}$ [1]	0.50 ± 0.31	0.22 ± 0.25	38
$Inh_{\text{Time RMS}}$ [3,4]	0.44 ± 0.30	0.19 ± 0.25	31
$RSA_{\text{R2 Std}}$ [1,3,4]	0.40 ± 0.28	0.57 ± 0.28	17
$RSP_{\text{Pks Mode}}$ [1]	0.61 ± 0.29	0.50 ± 0.30	08
$RSP_{\text{Rate Diff RMS}}$ [1,2,3]	0.33 ± 0.29	0.42 ± 0.29	06
RSP_{PSD3n} [1]	0.35 ± 0.29	0.43 ± 0.30	05
RSP_{PSD1n} [1]	0.48 ± 0.35	0.41 ± 0.31	04
$RR_{\text{HF gauss}}$ [1,3]	0.32 ± 0.23	0.68 ± 0.26	74
$RR_{\text{HF sband 3n}}$ [1]	0.47 ± 0.32	0.30 ± 0.28	15
$RR_{\text{Lorenz L2}}$ [2]	0.49 ± 0.30	0.35 ± 0.26	11
RR_{CVI} [2]	0.54 ± 0.29	0.42 ± 0.28	10
$PP_{\text{HF sband 5n}}$ [1]	0.23 ± 0.25	0.46 ± 0.30	28
PA_{RMS} [1,2]	0.53 ± 0.35	0.32 ± 0.27	20
$PA_{\text{Lorenz L}}$ [1]	0.44 ± 0.33	0.26 ± 0.25	17
PRS_{Mean} [2]	0.38 ± 0.35	0.55 ± 0.31	13
PP_{CSI} [1]	0.46 ± 0.29	0.33 ± 0.26	11
$PA_{\text{CSI modified}}$ [2]	0.40 ± 0.30	0.28 ± 0.27	09
PRT_{Median} [1,2,3]	0.44 ± 0.31	0.56 ± 0.31	08
$PTT_{\text{M Mode2}}$ [1]	0.50 ± 0.35	0.58 ± 0.28	05
PP_{Median} [1,2,3]	0.55 ± 0.31	0.47 ± 0.28	05
$PTT_{\text{M HF pond}}$ [1]	0.47 ± 0.28	0.54 ± 0.29	05
PRT_{LFp1oHF} [1]	0.38 ± 0.31	0.30 ± 0.27	05
PP_{Mode2} [1]	0.55 ± 0.33	0.49 ± 0.28	04
SKT_{Power} [1,2,3,4]	0.61 ± 0.35	0.37 ± 0.30	24
SKT_{Gradient} [1,2,3]	0.57 ± 0.29	0.38 ± 0.26	20

Selected feature with: [1]SVM-RFE, [2]LDA-RFE, [3]LR-RFE, and [4]XGB-RFE.

Besides, the CV-accuracy reported in Table III after RFE is delimited between 84 and 88%, except for DT. The CV-accuracy variability seems to be more dependent on the selected classifier (difference $> 4.5\%$) rather than the selected number of features (difference $< 3.5\%$). In fact, a linear SVM with an input of only five features can provide a reduced implementation complexity with a loss of only 1% of classification accuracy.

E. Classification improved with the SSM

Once we have selected the set of features (i.e., 25 features with RFE-SVM) and a linear SVM as classification method, we tested the subject-specific approach contribution compared to a general model (i.e., SSM vs. UBM). First, we trained the models as described in Section VI. The regularization term $C = 0.1$ of the UBM was selected based on a 5-fold CV on the training set. For the SSM, we selected $C = 0.001$ and $C_S = 0.1$ being the most common regularization terms found with a 5-fold CV on the training (data of Study 1).

TABLE V
PERFORMANCE OF THE UNIVERSAL BACKGROUND MODEL (UBM) VS. SUBJECT-SPECIFIC MODEL (SSM) ON AN UNSEEN DATA.

Study	Model	class	precision	recall	F1-score	samples
Study 1 Gamepad	UBM	B	0.81	0.76	0.79	123
		F3M	0.80	0.84	0.82	137
		avg	0.80	0.80	0.80	260
	SSM	B	0.89	0.83	0.86	123
		F3M	0.86	0.91	0.88	137
		avg	0.87	0.87	0.87	260
Study 2 FlyJacket	UBM	B	0.87	0.93	0.90	29
		F3M	0.92	0.86	0.89	28
		avg	0.90	0.89	0.89	57
	SSM	B	0.88	0.97	0.92	29
		F3M	0.96	0.86	0.91	28
		avg	0.92	0.91	0.91	57

Table V reports the comparison between UBM and SSM, tested on an unseen test set emulating an online CWM. The average accuracy of the UBM is 80.4%, and it is improved to 87.3% by the use of the SSM. The SSM shows a statistically significant improvement of the classification performance indicated by both the Wilcoxon rank-sum test [59] and the McNemar's test [60] over the 260 samples (p-value < 0.01). SSM improves the results for all the participants on CV, while one participant over 24 does not show the expected improvement on the final test set. This result may be explained by the need for more training data that could be used to better fit this participant's physiological response.

Furthermore, as shown in Table V, the higher performance of the SSM compared to the UBM is also confirmed on the test set acquired using FlyJacket (Study 2, Sec VII-C). In fact, the UBM reached a global accuracy of 89.5% that is improved to 91.2% using the SSM. However, the improvement (1 sample over 57) is not statistically significant, shown by both Wilcoxon rank-sum and McNemar's tests. For statistical results, additional data are needed. Nevertheless, a single misclassified sample in SAR missions can have a significant impact.

SSM obtains better performance than UBM because uses all the observations with a different weight. Those from other participants contribute to learn the general behaviour, with a regularization term C that allows a higher misclassification of such observations. Then, specific subject observations tune the margins between classes with a regularization term C_S to reinforce each specific subject. In light of the above, we can conclude that the personalized model performs in general better than the universal model.

Our results for the SSM are comparable with the state-of-the-art (See Table I), in particular with the work presented in [15], where the authors achieved an accuracy of 86%. Although with similar accuracy, our model is less complex and uses a reduced feature number. Another important difference is the test set selection, which was random in [15]. Instead, as a test set, we selected data from the last trial performed by

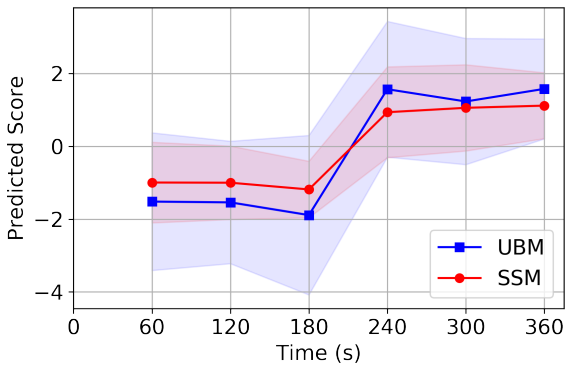

Fig. 9. Models performance comparison on a simulated online CWM (every 60s). The first 180 s correspond to B task followed by F3M task for other 180 s.

each subject, namely, Trial 3. For any classification problem that breaks the interchangeability hypothesis, such as the time-dependent CWM, a random training/test split should be avoided, as it yields a biased model evaluation. With a random split, the model learns from prospective data, commonly not available when designing and training a prediction model. Besides, the model is evaluated based on retrospective data, which are too similar to the training data. Hence, the classifier tends to look better than it is. Therefore, to estimate how well a model will work with new data, a time-dependent training/test split should be considered.

Also, the improved classification performance with FlyJacket vs. gamepad is assigned to the increased amount of training data. In the case of FlyJacket, the classifier weights were tuned based will all collected data from Study 1 and including Trial 1 of Study 2. Nevertheless, considering less training data (ignoring Day 2 of Study 1) reduces the accuracy of UBM from 86% to 82%.

Then, we assessed if our model using FlyJacket could suffer from possible movement artifacts, as they would differ from Task B to Task F3M. Thus, we minimized this risk because 93.5% of the samples used to train the classifier comes from Study 1, with the gamepad, in which the movements were minimal and limited to the thumbs. Moreover, all the features, normalization coefficients, and regularization terms were chosen using data only from Study 1. Thus, movement artifacts cannot significantly influence our classification results.

F. Emulated online cognitive workload monitoring

A visual representation of the emulated online CWM of both UBM and SSM is shown in Fig. 9. Since the order of the tasks was randomized, we only report the 76 samples of the sequences having consecutive transitions between B and F3M tasks. This analysis is based on Study 1 performed with the gamepad (Trial 3). During the first 180 seconds, participants performed the B task, a low workload level. For the last 180 seconds, participants performed the F3M task, a higher workload level. The detection was done on the test set, where features were extracted from a 60-second sliding window with no overlap. Negative and positive scores denote low and high workloads, respectively. A Wilcoxon rank-sum test with 76 samples indicates that the scores before and after 180 seconds are significantly different (p-value $< 10^{-8}$).

Another interesting aspect to note from Fig. 9 is the contradictory difference between the averaged predicted scores of the UBM and SSM. As the SSM is performing better than the UBM, we would expect to see a bigger absolute value of the SSM averaged score than the one of the UBM. However, the upper margin of the standard deviation of the predicted score reported in the interval between 60 and 180 seconds (Task B) and the lower margin in the interval between 240 and 360 seconds (Task F3M) seems similar for both UBM and SSM. This behaviour may be explained by the attempt of the SVM to choose the hyperplane that maximized the distance from it to the nearest data point on each side. Thus, as the SVM tends to maximize the margins, the SVM-based SSM performance may be limited to a consistent but marginal improvement.

Finally, comparing Fig. 5 and Fig. 9, we can see that both perceived and detected CWL are affected by a large variance. However, as shown in Fig. 9, such a variance is partially reduced using the SSM, which contributes better to fit the physiological response of a single subject.

IX. Conclusion

In this work, we have proposed a reliable subject-specific machine-learning algorithm for real-time CWM in SAR missions with drones. Our multimodal CWM model combines the information of features extracted from physiological signals (i.e., RSP, ECG, PPG, and SKT) noninvasively acquired. After an exhaustive investigation involving up to 384 features, we have selected only 25 required to get the highest classification accuracy. In addition, we have explored different feature normalization techniques to reduce both subject and day inter-variability, showing that a combination of day and subject normalization improves the detection accuracy.

Moreover, we have introduced a novel SVM based weighted-learning method suitable for subject-specific optimizations. With such a method, we distinguish between low and high CWL with an accuracy of 87.3%, on an unseen test set. Furthermore, we tested our model on ten new subjects using an advanced controller, reaching an average accuracy of 91.2%. Therefore, our model is valid to monitor CWL from rescuers piloting a drone with either traditional or advanced controllers.

The proposed methodology paves the way for detecting high levels of cognitive workload with sensors that can be included into a jacket. Our model can already operate in real-time to obtain information of the cognitive workload of the user. Such information can be used to improve shared-control systems by modulating the human-robot interaction and dynamically adapt the level of assistance, which will ensure an efficient execution of the missions. However, further investigations in real-life scenarios are needed to model other stressful conditions, which are not reproducible in laboratory tests. Moreover, there is a need to address a fine-grained detection in order to define a threshold for preventing a possible pilot's overload that could compromise the outcome of a search and rescue mission.

References

[1] I. Management Association, *Robotics: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, ser. Essential reference. IGI Global, 2013.

[2] J. Casper and R. R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2003.

[3] J. Y. C. Chen *et al.*, "Supervisory control of multiple robots: Human-performance issues and user-interface design," *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, vol. 41, no. 4, pp. 435–454, Jul 2011.

[4] K. H. Teigen, "Yerkes-dodson: A law for all seasons," *Theory & Psychology*, vol. 4, no. 4, pp. 525–547, Nov 1994.

[5] A. Marinescu *et al.*, "Exploring the relationship between mental workload, variation in performance and physiological parameters," *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 591–596, 2016.

[6] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, Jan 2002.

[7] S. G. Hart and L. E. Staveland, *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*, ser. Advances in Psychology. Elsevier, 1988, vol. 52, pp. 139–183.

[8] F. Dell'Agnola *et al.*, "Physiological characterization of need for assistance in rescue missions with drones," in *IEEE International Conference on Consumer Electronics (ICCE)*, 2018.

[9] H. Mansikka *et al.*, "Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 1, pp. 1–9, Feb 2019.

[10] B. Cain, "A Review of the Mental Workload Literature. Toronto," *Defence Research and Development Canada*, no. 1998, 2007.

[11] B. Ahmed *et al.*, "ReBreathe: A calibration protocol that improves stress/relax classification by relabeling deep breathing relaxation exercises," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 150–161, Apr 2016.

[12] M. Ranchet *et al.*, "Cognitive workload across the spectrum of cognitive impairments: A systematic review of physiological measures," *Neuroscience & Biobehavioral Reviews*, vol. 80, pp. 516–537, Sep 2017.

[13] J. Heard *et al.*, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, Oct 2018.

[14] G. Borghini *et al.*, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, Jul 2014.

[15] N. Momeni *et al.*, "Real-time cognitive workload monitoring based on machine learning using physiological signals in rescue missions," in *41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.

[16] F. Dell'Agnola *et al.*, "MBioTracker: Multimodal self-aware biomonitoring wearable system for online workload detection," *IEEE Transactions on Biomedical Circuits and Systems*, 2021.

[17] C. Rognon *et al.*, "Flyjacket: An upper body soft exoskeleton for immersive drone control," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2362–2369, Jul 2018.

[18] F. Dell'Agnola *et al.*, "Cognitive workload monitoring in virtual reality based rescue missions with drones," in *Int Conf Human-Computer Interact*, vol. 12190 LNCS. Copenhagen, Denmark: Springer, Cham, jul 2020, pp. 397–409.

[19] V. Montesinos *et al.*, "Multi-modal acute stress recognition using off-the-shelf wearable devices," in *41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.

[20] L.-l. Chen *et al.*, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, vol. 85, pp. 279–291, Nov 2017.

[21] E. T. Solovey *et al.*, "Classifying driver workload using physiological and driving performance data: two field studies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[22] D. Giakoumis *et al.*, "Subject-dependent biosignal features for increased accuracy in psychological stress detection," *International Journal of Human-Computer Studies*, vol. 71, no. 4, pp. 425–439, Apr 2013.

[23] A. Tjolleng *et al.*, "Classification of a driver's cognitive workload levels using artificial neural network on ecg signals," *Applied Ergonomics*, vol. 59, pp. 326–332, Mar 2017.

[24] M. Gjoreski *et al.*, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, pp. 159–170, Sep 2017.

[25] F. T. Eggemeier *et al.*, "Workload assessment in multi-task environments," in *Multiple Task Performance*, 1991.

[26] J. C. Christensen *et al.*, "The effects of day-to-day variability of physiological data on operator functional state classification," *NeuroImage*, vol. 59, no. 1, pp. 57–63, 2012, neuroergonomics: The human brain in action and at work.

[27] T. Luong *et al.*, "Towards real-time recognition of users mental workload using integrated physiological sensors into a vr hmd," in *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2020, pp. 425–437.

[28] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, Jun 2005.

[29] T. Heine *et al.*, "Electrocardiographic features for the measurement of drivers' mental workload," *Applied Ergonomics*, vol. 61, pp. 31–43, May 2017.

[30] D. Novak *et al.*, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interacting with Computers*, vol. 24, no. 3, pp. 154–172, May 2012.

[31] L. Han *et al.*, "Detecting work-related stress with a wearable device," *Computers in Industry*, vol. 90, pp. 42–49, Sep 2017.

[32] R. J. Lysaght *et al.*, "Operator workload: Comprehensive review and evaluation of operator workload methodologies," *United States Army Research Institute for the Behavioral Sciences, Technical Report*, 1989.

[33] D. Carneiro *et al.*, "New methods for stress assessment and monitoring at the workplace," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 237–254, Apr 2019.

[34] D. Novak *et al.*, "Workload estimation in physical human-robot interaction using physiological measurements," *Interacting with Computers*, vol. 27, no. 6, 2015.

[35] B. Cinaz *et al.*, "Monitoring of mental workload levels during an everyday life office-work scenario," *Personal and Ubiquitous Computing*, vol. 17, no. 2, pp. 229–239, Feb 2013.

[36] P.-K. Jao *et al.*, "EEG correlates of difficulty levels in dynamical transitions of simulated flying and mapping tasks," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 2, pp. 99–108, 2020.

[37] A. Arza *et al.*, "Measuring acute stress response through physiological signals: towards a quantitative assessment of stress," *Medical and Biological Engineering and Computing*, pp. 1–17, Aug 2018.

[38] M. van Dooren *et al.*, "Emotional sweating across the body: comparing 16 different skin conductance measurement locations." *Physiology & behavior*, vol. 106, no. 2, pp. 298–304, may 2012.

[39] J. Pan and W. J. Tompkins, "A Reat-Time QRS Detection Algorithm," *IEEE transactions on bio-medical engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[40] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, Mar 1996.

[41] W. H. Press and G. B. Rybicki, "Fast Algorithm for Spectral Analysis of Unevenly Sampled Data," *Astrophysical Journal, Part 1*, vol. 338, pp. 277–280, 1989.

[42] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, no. 258, Sep 2017.

[43] S. W. Porges, "Cardiac vagal tone: A physiological index of stress," *Neuroscience & Biobehavioral Reviews*, vol. 19, no. 2, pp. 225–233, Jun 1995.

[44] P. Grossman and E. W. Taylor, "Toward understanding respiratory sinus arrhythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions," *Biological Psychology*, vol. 74, no. 2, pp. 263–285, Feb 2007.

[45] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between 2 methods of clinical measurement," *Lancet*, vol. 8476, pp. 307–310, 1986.

[46] W. Zhang *et al.*, "Task-specific normalization for continual learning of blind image quality models," 2021.

[47] S. Theodoridis *et al.*, *Introduction to Pattern Recognition*. Boston: Academic Press, 2010.

[48] H. Akoglu, "User's guide to correlation coefficients," *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[49] I. Guyon *et al.*, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[50] D. Reynolds, *Universal Background Models*. Boston, MA: Springer US, 2009, pp. 1349–1352.

[51] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, Jan 2006.

[52] MATLAB, *R2016a*. Natick, Massachusetts: The MathWorks Inc., 2016.

[53] D. Goepfert *et al.*, *Reglement Einsatzführung*. Bern: Feuerwehr Koordination Schweiz FKS, 2015.

[54] A. Mondal *et al.*, "A noise reduction technique based on nonlinear kernel function for heart sound analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 775–784, 2018.

[55] J. Martín *et al.*, "On the regressand noise problem: Model robustness and synergy with regression-adapted noise filters," *IEEE Access*, vol. 9, pp. 145 800–145 816, 2021.

[56] S. Ansari *et al.*, "Motion artifact suppression in impedance pneumography signal for portable monitoring of respiration: An adaptive approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 2, pp. 387–398, 2017.

[57] Q. Zhang *et al.*, "Motion artifact removal for ppg signals based on accurate fundamental frequency estimation and notch filtering," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2965–2968.

[58] J. Miehlbradt *et al.*, "Data-driven body-machine interface for the accurate control of drones," *Proceedings of the National Academy of Sciences*, vol. 115, no. 31, pp. 7913–7918, Jul 2018.

[59] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, no. 6, pp. 80–83, 1945.

[60] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.