Real-Time Cognitive Workload Monitoring Based on Machine Learning Using Physiological Signals in Rescue Missions*

Niloofar Momeni, Fabio Dell'Agnola, Adriana Arza, and David Atienza[1]

Abstract— **High levels of cognitive workload decreases human's performance and leads to failures with catastrophic outcomes in risky missions. Today, reliable cognitive workload detection presents a common major challenge, since the workload is not directly observable. However, cognitive workload affects several physiological signals that can be measured non-invasively. The main goal of this work is to develop a reliable machine learning algorithm to identify the cognitive workload induced during rescue missions, which is evaluated through drone control simulation experiments. In addition, we aim to minimize the computing resources usage while maximizing the cognitive workload detection accuracy for a reliable real-time operation. We perform an experiment in which 24 subjects played a rescue mission simulator while respiration, electrocardiogram, photoplethysmogram, and skin temperature signals were measured. State-of-the-art feature-based machine learning algorithms are investigated for cognitive workload characterization using learning curves, data augmentation, and cross-validation techniques. The best classification algorithm is selected, optimized, and the most informative features are selected. Finally, the generalization power of the optimized model is evaluated on an unseen test set. We obtain an accuracy level of 86% on the new unseen datasets using the proposed and optimized eXtreme Gradient Boosting (XGB) algorithm. Then, we reduce the complexity of the machine learning model for future implementation on resource-constrained wearable embedded systems, by optimizing the model and selecting the 26 most important features. Overall, a generalizable and low-complexity machine learning model for cognitive workload detection based on physiological signals is presented for the first time in the literature.**

Index Terms— **Cognitive workload, stress, physiological signals, machine learning, XGBoost, rescue missions.**

I. INTRODUCTION

Increasing difficulty of different tasks imposes varying levels of cognitive workload depending on the human's capacities and skills [1]. High levels of cognitive workload leads to failures with catastrophic outcomes such as accident [2], since it significantly decreases human's performance [3]. Therefore, detecting the high cognitive workload can improve working conditions and in general the quality of life in our society.

Cognitive workload detection in high risky tasks has received special attention. One recent application is in rescue missions with drones to characterize need for assistance [4]. When a disaster occurs, the rescuer has to handle many complex activities involving cognitive workload such as monitoring the pathway, controlling the drone, searching for victims, and making a proper decision to manage the damaged situation [4]. Therefore, detecting the excessive cognitive workload induced during flying a drone is important for preventing accidents and hazards.

An excessive cognitive workload can create very high stress state [5]. Stress is often recognized as the response of the human body to adapt to external demands when, as defined in [6], an important situation taxes or exceeds his or her capabilities and resources. Both stress and cognitive workload concepts involve environmental demands and the ability of the person, or rescue mission operators in particular, to cope with those demands, although these two concepts come from different theoretical backgrounds [7].

Nowadays, a reliable detection of both workload and stress presents a common major challenge, since they are not directly observable. However, there are several approaches to assess cognitive workload and stress, i.e., subjective questionnaires, performance analysis, and physiological reaction analysis [7]. Subjective questionnaires cannot be used frequently and makes it unsuitable for a continuous workload monitoring [8]. Performance analysis typically measures the difference between the expected and the actual performance [1], but it is not possible to measure the performance online. However, physiological reaction analysis can be used for a reliable, non-intrusively, multi-modal (considering multiple signals), and in a real-time monitoring of the cognitive workload [4], [8]–[12].

Recently, state-of-the-art studies that target classification of cognitive workload and stress based on physiological signals started to employ machine learning techniques [10]. However, to the best of our knowledge, no one has evaluated so far the model's generalization on unseen datasets. One of the first relevant studies is presented by Healey and Picard [13]. They distinguish levels of driver stress with an accuracy of 97%, on the training set, with linear discriminant analysis (LDA) algorithm across 24 drivers using electrocardiogram (ECG), electrodermal activity (EDA), and respiration (RSP) signals analysis. Recently, Chen et al. [11] use the same database as in [13] to address the stress status but for 14 subjects and considering Support Vector Machine (SVM) machine learning algorithm for workload detection. They reach an accuracy of over 99% on the training set and 89% in cross-validation. Both works only use training set or cross-validation set to assess their accuracy without evaluating its generalization on unseen (or real-life testing) datasets.

*This work has been partially supported by the NCCR Robotics through the Symbiotic Drone project, and by the ONR-G through the Award Grant No. N62909-17-1-2006.

[1]N. Momeni, F. Dell'Agnola, A. Arza, and D. Atienza are with the Embedded Systems Laboratory of Swiss Federal Institute of Technology Lausanne, Switzerland. {niloofar.momeni, fabio.dellagnola, adriana.arza, david.atienza} @ epfl.ch

Similarly, Gjoreski et al. [14] evaluate their model using only a leave-one-out cross-validation. In this case, they obtain a 73% accuracy detecting stress of daily life activity. They use an SVM algorithm based on data collected in 55 days from photoplethysmogram (PPG), skin temperature (SKT), and EDA signal as well as from a 3-axis accelerometer. In fact, most of the state-of-the-art studies evaluate their methodology on cross-validation sets [8], [10]–[13], which is not a reliable technique to evaluate the generalization power in machine learning based approaches [15].

The tasks used to induce cognitive workload and stress is another factor to analyze when comparing those approaches. The stronger the stressor agent is, like in real driving task [8], [11], the better accuracy results are reported. When the stimuli of the stressor is not very strong, the classification problem gets harder since more difficult it will be to learn from the dataset. Therefore, in those cases, we need to consider an advance machine learning algorithm combined with a multi-modal analysis that fuses the information in several physiological signals.

As a result, the main goal of this project is to develop a reliable machine learning algorithm to identify the cognitive workload induced during rescue missions with drones simulation experiments. The proposed multi-modal cognitive workload detection model combines the information in features extracted from the ECG, RSP, PPG and SKT signals.

Additionally, we aim for reducing the complexity of our machine learning algorithm and optimizing its resource usage for real-time operation. This reduction is particularly important in the context of wearable embedded systems, which are highly constrained in terms of computational capacity, memory storage, and battery lifetime Our main contributions are as follows:

- A new and reliable machine learning model for cognitive workload detection based on multimodal physiological signals that is generalizable.
- Minimization of the resource usage and complexity of the machine learning model by selecting only 26 of the most important features and optimizing the model's hyper-parameters.
- Obtaining an accuracy of 86% on the new unseen datasets, which is the first assessment of this kind in the literature for cognitive workload detection in rescue missions based on physiological signals. Moreover, we reach an accuracy of 94.8% on cross-validation sets, which is higher than the latest state-of-the-art studies.

II. Cognitive Workload Classification Methodology

A. Experiment Protocol and Setup

The developed simulation environment presented in [4] is used in this work to track the workload influence on search and rescue missions with drones. In this experimental setup, the bio-signals are recorded on users while using the simulator according to the following scenarios: a baseline state (physiological condition before starting the flight tasks, B), a waypoint following task (flying task, F), and flying

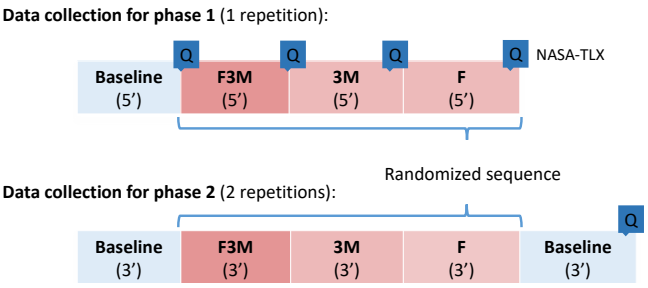

Fig. 1. Signal acquisition protocol. Q: Questionnaire in resting period [4].

while tagging three objects of interest simultaneously (F3M task). This developed scenario is quite similar to a real search and rescue mission, where a rescuer has to control a drone and map the environment identifying the damaged situation.

The data we use in this work was collected from 24 subjects, over two different days. The participants performed three trials per day. Accordingly, the experiment consists of data collection in two phases as in Figure 1. In the first trail the main tasks are divided by a resting period, while in trials 2 and 3, the sequence of tasks has no resting periods in between. The cognitive workload level is reported by the subjects after each task, based on the National Aeronautics and Space Administration Task Load Index (NASA-TLX) [16]. Data are labeled based on the rating of stress levels from the questionnaire as high and low workload , i.e. values higher than 50% and lower than 12% respectively.

This study is part of the approved protocol number PB_2017-00295 granted by the Cantonal Ethics Commissions for Human Research Vaud and Geneva. During the experiments we recorded all the physiological signals using the Biopac MP160 data acquisition system [17].

B. Cognitive Workload Feature Extraction

In the gathered data we extract several biomarkers based on the type of the aforementioned physiological signals and the latest state-of-the-art results [4], [18]. These biomarkers are segmented in 60-second-length sliding windows. Different sliding intervals are used, i.e 60, 50, 40, 30, 20, and 10 seconds to assess the effect of the data overlaping in the classification performance. Then, for each window, a set of features is obtained using time-domain and frequency-domain analysis. The considered biomarkers from each bio-signals are the following:

- ECG: The RR internals are extracted from the ECG signals as presented in Figure 2. From the RR interval, several time and frequency domain bio-markers are extracted based on the Heart Rate Variability (HRV) analysis of the RR interval series [19]. Non-linear features are also extracted from Poincaré plot indicating vagal and sympathetic function, as reviewed in [20]. They are the following: the length of the transverse axis ($SD1$), which is vertical to the line $\mathrm{NN}_k = \mathrm{NN}_{k+1}$; the length of the longitudinal axis ($SD2$), which is parallel with the line $\mathrm{NN}_k = \mathrm{NN}_{k+1}$; the ratio $SD2/SD1$,

called Cardiac Sympathetic Index (CSI); the modified CSI ($SD2^2/SD1$); and the $log_{10}(SD2 \cdot SD1)$, which is also called Cardiac Vagal Index (CVI) [19].

- PPG: Several bio-markers are computed from the pulse wave of the PPG signal, which are represented in Figure 2. They are pulse period (PP), pulse transit time (PTT), pulse wave rising time (PRT), pulse wave decreasing time (PDT), pulse width until reflected wave (PWR), and pulse width (PW). From each PPG bio-marker, time and frequency domain heart rate variability (HRV) features, as well as Poincaré plot indexes features are extracted, as aforementioned for the ECG features.

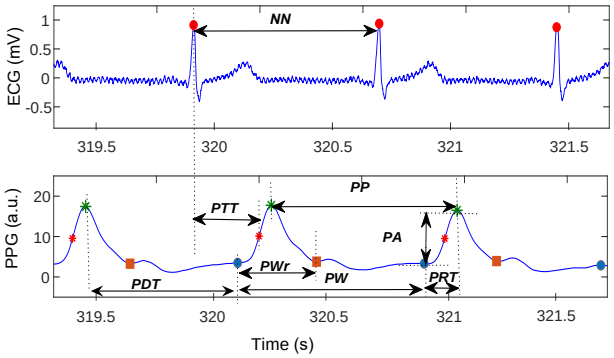

Fig. 2. Bio-markers extracted from ECG and PPG signals [18]

- RSP: Several common bio-markers of the respiration signal are computed and statistical features are extracted. In particular, we include respiration rate (F_R), time and frequency analysis of respiration period, volume of air inhaled or exhaled in one minute (Minute Ventilation), ratio of inhalation to exhalation duration (IE Ratio), etc.
- SKT: Two features are obtained from the filtered skin temperature signal: the ΔT that is computed as the difference between the last (not missing) value of the window minus the first one and the T_{Pt}, as in [18].

Overall, 385 features are extracted and evaluated from the considered physiological signals.

C. Cognitive Workload Classification

We look at a general machine learning algorithm that can reliably detect elevated cognitive workload across individuals. We consider only the low and high cognitive workload, which correspond to baseline and F3M tasks, respectively. The block diagram in Figure 3 shows the methodology steps considered in this study to obtain a valid model for cognitive workload classification.

Several state-of-the-art machine learning algorithms are trained, cross validated, and the best one is selected. In addition, we consider a wide variety of physiological features to investigate the contribution of several different biomarkers for detection of the cognitive workload in the initial models. Then, the most informative features are selected by removing the low importance ones. Next, model's complexity is controlled by optimizing its hyper-parameters. Finally, the generalization power of the optimized model is evaluated on an unseen test set.

1) Cognitive Workload Classification Algorithm Selection: We explore the best machine learning algorithm in the context of our cognitive workload detection problem. According to the "No Free Lunch" theorem [21], there is no one predictive model that performs best for every problem. Thus, different classifiers can well suit for different problems, depending on many factors, such as, the problem type, data size, and data structure [22].

Seven feature-based machine learning algorithms are considered in this analysis: Logistic Regression (LogReg), Decision Trees Classifier (DTC), k Nearest Neighbor (k-NN), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), SVM, and eXtreme Gradient Boosting (XGB).

We compare the performance of these algorithms versus: (1) the training data size, using a 10-fold cross-validation learning curve, and (2) the data augmentation, using an 8-fold cross-validation score curve over the overlap factor.

Cross-Validation Learning Curve Analysis: A 10-fold cross-validation to get smooth mean validation and train score curves is implemented for the seven classification algorithms, each time with 20% randomly selected data as a validation set. We observe the validation and training score of each estimator by varying the number of training samples on learning curve [22]. This learning curve is one of the main tools that enables us to find out how much we benefit from adding more training data and whether the estimator suffers more from a variance error or a bias error.

Data Augmentation with a Cross-Validation Analysis: Data augmentation can make a classification algorithm more robust. It is important however to note that only informative new samples can improve our model's performance and not all of the artificially generated data are informative. There are several ways for data augmentation, including adding noise to signals, changing the window length while segmenting signals, and segmenting signals with a certain degree of overlap. In this study, we artificially generate more data using overlap. The overlap factor determines how much information is shared among successive windows. To investigate overlap effect on the classification performance of different algorithms, we implement an 8-fold cross-validation on dataset with sliding windows of 60, 50, 40, 30, 20, and 10 seconds, which represent overlap percentages of 0%, 16%, 33%, 50%, 66% and 83%.

The training and cross-validation sets should not have any overlap to avoid having bias in evaluation. Therefore, we divide the data into training and validation sets based on the subjects, trials, and the days of the experiments. In total, we have an 8-fold cross-validation approach applied on each of the six datasets since the data have been collected in two days, two trials, and odd or even subject's number.

2) Cognitive Workload Feature Selection: Our dataset of size 670×385 suffers from high dimensionality. Having a large feature set makes the model more complex and increases the chance of overfitting. As a result, the model becomes more sensitive to error due to variance. Moreover,

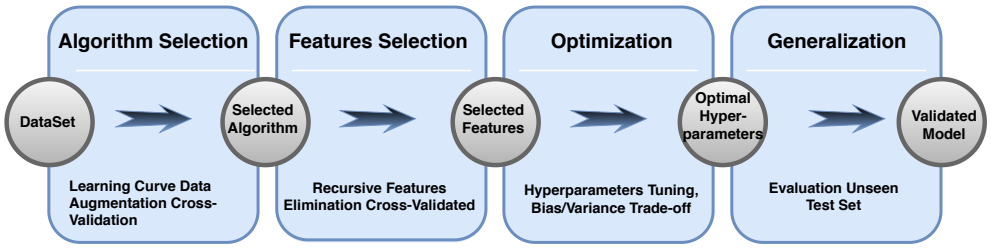

Fig. 3. Block diagram showing the steps to obtain a valid model for cognitive workload classification.

large amount of features leads to increase computational time, power consumption, and elevates cost when the model is deployed to embedded systems [23]–[26]. Therefore, reducing the number of features (without much loss of the total information to suppress model's complexity) is fully recommendable [27].

Feature selection [22] is a suitable method to remove worst performing features by eliminating the irrelevant (features with low importance) and redundant (highly correlated) ones. In addition, Recursive Feature Elimination (RFE) is a feature selection method prposed in [27], which recursively eliminates the least important features in a loop without loosing classification performance.

In particular, Recursive Feature Elimination with Cross-Validation (RFECV) selects the optimal number of features using RFE based on the cross-validation score for a given estimator [27]. Therefore, in order to obtain the optimal number of features, a shuffle-split cross-validation with 10 iteration and validation size of 33% is used with RFE by taking the selected classifier from the previous section. The selected number of features is observable from the trend of the RFECV curve. The curve arrives to an excellent accuracy and enters in a relatively steady state when the most informative features are captured. Finally, after applying the RFECV, a new dataset with a selected subset of features is considered for the subsequent analysis using hyper-parameter optimization.

3) Hyper-parameter Optimization: In machine learning, hyper-parameters represent higher level properties of a model such as complexity, or how fast a model can learn from a dataset [28]. The value of a hyper-parameter has a significant effect on the result of model's performance. Hyper-parameter optimization increases the accuracy score of a model, reduces the complexity and variance of an algorithm, and is useful in the context of embedded systems, which are limited in terms of resources. Thus, we explore the optimal values of the hyper-parameters of the selected classifier for the best subset of features on the best augmented dataset.

First, depending on the selected model, we choose two hyper-parameters that have more effect on the result of the model's performance. Then, we select a suitable range for each of the hyper-parameters depending on their usual space range. Next, we train the selected model multiple times using cross-validation for different values of each of the hyper-parameter's range. Next, we plot a 3D curve representing the average value of the model's cross-validation score for different values of hyper-parameters. We select the optimal values that lead to the maximum cross-validation score for both hyper-parameters. Then, we evaluate those values using validation curves.

4) Generalization: The learned model must be able to characterize cognitive workload on *new, previously unseen* inputs, not just those ones which our model was trained. The ability of the learned model to fit on unseen data or test set is called generalization [15].

Although we can optimize the algorithm using validation set, it is not sufficient to evaluate model's generalization power based on validation score because this evaluation is biased and it not a good representitive of the generalization. In this study, the dataset from the second repetition in phase two (i.e., trial two) is considered as the test set. The factors that determines the generalization power of an algorithm are (1) the cross-validation score, (e.g., accuracy should be high) and (2) the gap between cross-validation and test score should be small.

III. Experimental Results

A. Cognitive Workload Classification Algorithm Selection

We investigate the performance of the LogReg, DTC, k-NN, LDA, GNB, SVM, and XGB algorithms versus: (1) the training data size, using a 10-fold cross-validation learning curve, and (2) the data augmentation, using an 8-fold cross-validation score curve over the overlap factor. Since the dataset is balanced, accuracy is considered a valid metric for performance evaluation.

1) Cross-Validation Learning Curve Analysis: We execute training runs for the seven classification algorithms with 10-fold cross-validation. We observe the validation and training scores of each estimator by varying the number of training samples on the learning curve.

The result of the learning curve analysis for seven classifiers is shown in Figure 4. The learning curve of the k-NN, LDA, and GNB algorithms do not converge to a plateau and their curve trend is not flat. Thus, these algorithms need more data to become stable. Moreover, the LDA, DTC, and (to some extent) XGB algorithms suffer from high variance since there exists a big gap between training and validation scores and thus, these models are considered complex. Note that a complex model does not necessarily lead to overfitting if the validation accuracy is high. The LDA and DTC models are overfitting because their validation score is low. Although the XGB algorithm is a complex model, due to the high

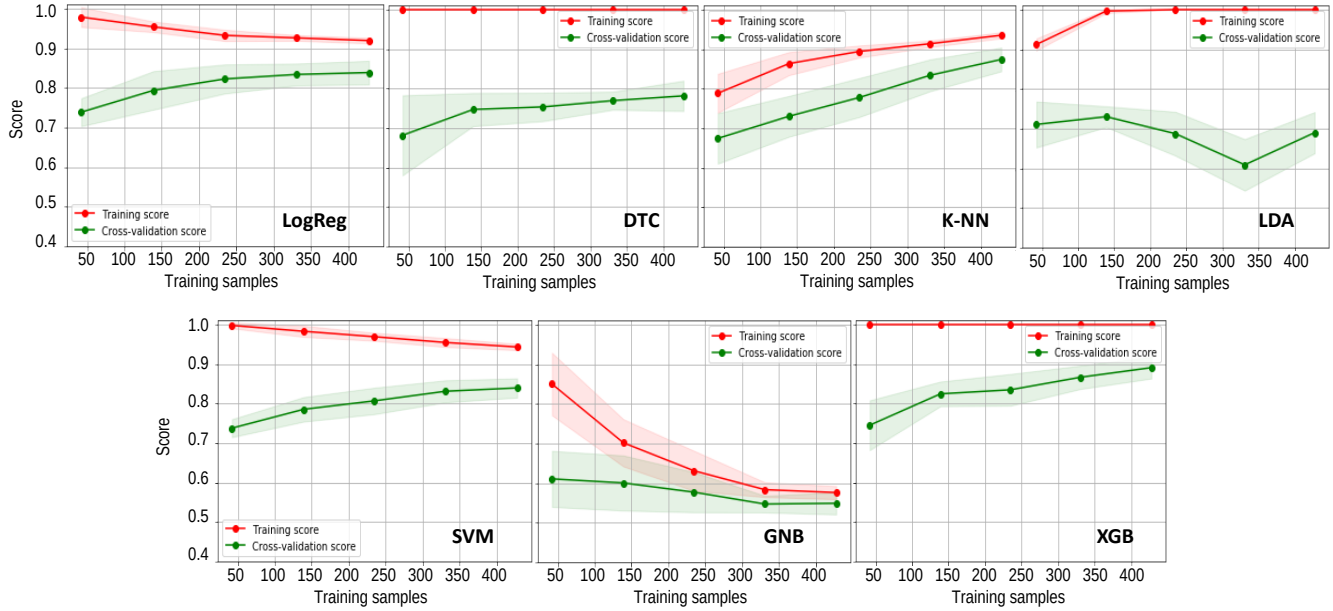

Fig. 4. Learning curves for different classification algorithms.

variance, but it still has a high score in the validation set. In addition, the GNB algorithm has a high bias since both of the training and validation scores converge to a low value. The results of learning curve experiment analysis are summarized in Table I.

TABLE I

PERFORMANCE COMPARISON OF DIFFERENT CLASSIFICATION ALGORITHMS.

Algorithm	Training Acc.	Val Acc. (std)	Stable	Low Variance	Low Bias
LogReg	92%	84% (0.03)	✓	✓	✓
DTC	100%	79% (0.04)	✓		✓
K-NN	94%	87%(0.03)		✓	✓
LDA	100%	69% (0.05)			✓
GNB	57%	55% (0.03)		✓	
SVM	93%	84% (0.02)	✓	✓	✓
XGB	100%	89% (0.03)	✓		✓

The accuracy on the validation set of the XGB classifier is the highest and after that the k-NN, SVM, and LogReg algorithms have high accuracy. The LDA and GNB do not have a good discrimination ability and they have the lowest performance among other algorithms.

2) Data Augmentation with Cross-Validation Analysis: The resulting learning curves in Figure 4 show the possibility of model's performance improvement by increasing the amount of training data. However, the limited training data is one of the main constraining factors to achieve better classification performance in this type of studies. Therefore, in this work we artificially generate more data using data augmentation techniques. Our results for data augmentation with an 8-fold cross-validation experiment is shown in Figure 5 and is summarized in Table II. Comparing the performance of the classifiers, the XGB algorithm has the best performance among the others. The dataset with 50% overlap factor resulted in the maximum cross-validation accuracy of 86% on the XGB algorithm.

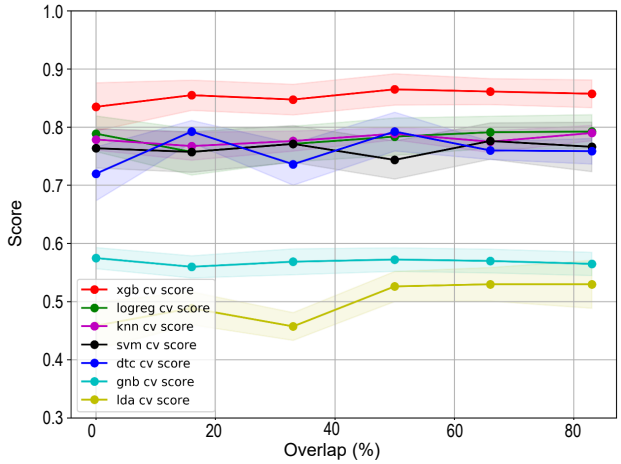

Fig. 5. Data augmentation with an 8-fold cross-validation analysis

Finally, based on both results of the learning curve and data augmentation analysis, the XGB algorithm with 50% overlap dataset is the best algorithm targetting our cognitive workload classification problem. Moreover, XGB does not suffer from high bias and its cross-validation accuracy is the highest. Additionally, the result is reliable since XGB's learning curve tends to converge to a plateau, which means that it does not need more training instances to become stable. The XGB algorithm is an advance machine learning algorithm that was developed by Tianqi Chen in 2016 [29] and until now, it has not been investigated within any of the studies in the context of cognitive workload classification.

Although the XGB suffers from a high variance, it is possible to reduce the variance by removing features with low importance, as well as tuning the parameters of the XGB model, as we have done in this work.

TABLE II

DATA AUGMENTATION RESULTS WITH CROSS-VALIDATION.

Algorithm	Acc. 0% Aug.	Max. Acc.	%Aug. for Max Acc.
LogReg	79%	79%	0%
DTC	72%	79%	16%
K-NN	78%	79%	50%
LDA	46%	53%	66%
GNB	58%	58%	0%
SVM	76%	78%	66%
XGB	83%	86%	50%

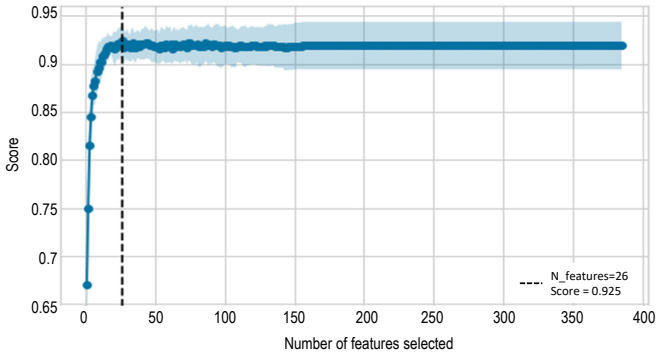

Fig. 6. Recursive feature elimination with cross-validation curve.

B. Cognitive Workload Feature Selection

We implement the RFECV method on our dataset to obtain the optimal number of features.Since we have a data-set with a large number of features, we remove two features in each step. Figure 6 illustrates that the classification accuracy is improved with the increasing number of selected features. After about 26 features the classification accuracy enters in a relatively steady state. Thus, the selected optimal number of features is 26. However, we achieve a cross-validation accuracy of 92.5%. This result implies that we do not lose information by removing the 359 features. Consequently, the complexity of our XGB model is reduced.

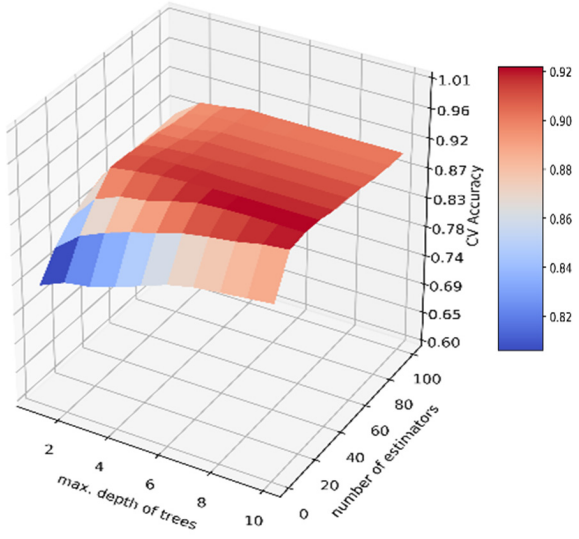

Fig. 7. Coarse-grain in two-dimensional XGB hyper-parameter

C. Our proposed of XGB Optimization

To avoid falling in a local optimum, we first perform a coarse-grained two-dimensional hyper-parameter sweeping to identify the promising region in the space. We show the mean cross-validation accuracy versus the *n_estimators* and the *max_depth* of the trees in Figure 7. Then, we pick the coordinate corresponding to the maximum of the mean cross-validation accuracy, i.e., 100 for the *n_estimators* and 4 for the *max_depth* of the trees, which leads to a mean cross-validation accuracy of 94.8%.

After finding a promising region in the space of hyper-parameters, we evaluate the model's performance with respect to the bias vs. variance trade-off using validation curves. As a result, the required trade-off for our XGB model is met at *max_depth*=4 and *n_estimators*=80 with cross validation accuracy of 94.8%.

D. XGB Model Generalization

In this section we evaluate the proposed XGB model on the new unseen test set.The results are presented in Table III. As this table shows, the XGB algorithm achieves an accuracy of 82% on the unseen test set using only the original dataset. Moreover, with data augmentation technique, we could increase the accuracy of the test set from 82% to 84%. The accuracy on the test set after applying the RFECV is 84%. Although we eliminate 359 features, there is no performance loss. Hence, we significantly reduce the complexity of the model. In addition, we feed the resulted optimized hyper-parameters into the XGB model, which improves the performance from 84% to 86% on the test set. Beside increasing the performance, we increase the training and inference efficiency and decrease the complexity of the model by reducing the number of trees from 100 to 80 in our XGB model.

In addition, the results in Table III highlight that the cross-validation accuracy in every step is high and the gap between cross-validation and test accuracy is small. Therefore, we can conclude that the generalization power of our optimized XGB model is high. Indeed, this model performs well on classifying between "low" and "high" cognitive workload levels.

Overall, although several studies exist in the literature that employ machine learning algorithms for workload and stress detection based on physiological signals, neither of them evaluates and reports classification results on a new unseen dataset as we perform in this work for the first time.. Infact, state-of-the-art studies evaluate their methodology on the cross-validation set, which is not reliable enough to evaluate the generalization power in machine learning. Even considering only the cross-validation technique, we obtain higher accuracy level of 94.8% than state-of-the-art cognitive and stress classification studies based on multimodal physiological signals. Finally, using a careful literature analysis, the previous best results of cross-validation accuracy for cognitive workload detection based on multimodal physiological signals reported by [8], [11], [12], [30] are 94%, 90%, 89%

TABLE III

CLASSIFICATION RESULTS EVALUATION USING UNSEEN TEST SET.

Description	Aug.	RFECV	Optimization	Val. Acc.	Test Acc.
Original data, 385 features, default XGB				83%	82%
50% Aug. data, 385 features, default XGB	✓			86%	84%
50% Aug. data, 26 features, default XGB	✓	✓		92%	84%
50% Aug. data, 26 features, tuned XGB	✓	✓	✓	94.8%	86%

and 86%, respectively. These figures, as the results of this section have shown, are worse than those obtained with the new proposed XGB model.

IV. CONCLUSION

In this work we have proposed a new reliable machine learning algorithm to identify the cognitive workload in rescue missions with drones. We have performed an experiment with 24 subjects playing a rescue mission simulator. We have evaluated several machine learning algorithms to investigate their ability to predict the cognitive workload on our dataset. The XGB algorithm demonstrates the highest performance in cross-validation. Our multi-modal cognitive workload detection model combines the information in features extracted from several physiological signals. We have reached an accuracy at the level of 86% on the new unseen test sets, presented for the first time in the literature and an accuracy level of 94.8% on cross-validation set, which is higher than state-of-the-art studies. In addition to the successful classification of the cognitive workload with high accuracy using the XGB, significant improvements are obtained in terms of inference and model complexity, for future implementation on the resource-constrained embedded systems, by extracting the most important features and optimizing the learning model. We have shown that the number features for cognitive workload classification can be as small as 26 features, without any major performance loss.

REFERENCES

[1] D. Gopher and E. Donchin, "Workload: An examination of the concept." *Handb Percept Hum Perform*, jan 1986.

[2] G. F. Wilson, "An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures," *Int J Aviat Psychol*, vol. 12, no. 1, 2002.

[3] N. Moray, *Mental workload : its theory and measurement*. Published in coordination with NATO Scientific Affairs, Plenum Press, 1979.

[4] F. Dell'Agnola, L. Cammoun, and D. Atienza, "Physiological characterization of need for assistance in rescue missions with drones," in *IEEE Int Conf Consum Electron*, 2018.

[5] H. Selye, "Stress and the general adaptation syndrome," *Br Med J*, vol. 1, no. 4667, 1950.

[6] R. F. Baumeister and K. D. Vohs, "Self-Regulation, ego depletion, and motivation," *Soc Personal Psychol Compass*, vol. 1, no. 1, 2007.

[7] B. Cain, "A Review of the Mental Workload Literature," *Def Res Dev Toronto*, 2007.

[8] E. T. Solovey, M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler, "Classifying driver workload using physiological and driving performance data: two field studies," in *Proc SIGCHI Conf Hum Factors Comput Syst*. ACM, 2014.

[9] B. Mehler, B. Reimer, J. F. Coughlin, and J. A. Dusek, "Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers," *Transp Res Rec J Transp Res Board*, vol. 2138, no. 1, 2009.

[10] J. Heard, C. E. Harriott, and J. A. Adams, "A Survey of Workload Assessment Algorithms," *IEEE Trans Human-Machine Syst*, 2017.

[11] L.-l. lan Chen, Y. Zhao, P.-f. fei Ye, J. Zhang, and J.-z. zhong Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Syst Appl*, vol. 85, 2017.

[12] S. Betti, R. M. Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. D. Ry, and F. Cavallo, "Evaluation of an Integrated System of Wearable Physiological Sensors for Stress Monitoring in Working Environments by Using Biological Markers," *IEEE Trans Biomed Eng*, vol. 65, no. 8, aug 2018.

[13] J. A. Healey and R. W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," *IEEE Trans Intell Transp Syst*, vol. 6, no. 2, 2005.

[14] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J Biomed Inform*, vol. 73, 2017.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[16] S. G. Hart, L. E. Staveland, S. G. Hart, and L. E. Stavenland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Adv Psychol*. North-Holland, jan 1988, vol. 52.

[17] Biopac, "MP160 Data Acquisition Systems." [Online]. Available: https://www.biopac.com/product/mp150-data-acquisition-systems/

[18] A. Arza, J. M. Garzón-Rey, J. Lázaro, E. Gil, R. Lopez-Anton, C. de la Camara, P. Laguna, R. Bailon, and J. Aguiló, "Measuring acute stress response through physiological signals: towards a quantitative assessment of stress," *Med Biol Eng Comput*, 2018.

[19] 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, mar 1996.

[20] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *J Biomed Inform*, vol. 59, no. C, feb 2016.

[21] D. H. Wolpert and W. G. Macready, "No Free Lunch Theorems for Optimization," *Trans Evol Comp*, vol. 1, no. 1, apr 1997.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in {P}ython," *J Mach Learn Res*, vol. 12, 2011.

[23] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-Time Event-Driven Classification Technique for Early Detection and Prevention of Myocardial Infarction on Wearable Systems," *IEEE Trans Biomed Circuits Syst*, vol. 12, no. 5, oct 2018.

[24] F. Forooghifar, A. Aminifar, and D. Atienza, "Self-Aware Wearable Systems in Epileptic Seizure Detection," *Euromicro Conf Digit Syst Des*, p. 7, 2018.

[25] G. Surrel, A. Aminifar, F. Rincón, S. Murali, and D. Atienza, "Online obstructive sleep apnea detection on medical wearable sensors," *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 4, pp. 762–773, 2018.

[26] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017, pp. 1–4.

[27] B. Bengfort, N. Danielsen, R. Bilbro, L. Gray, K. McIntyre, G. Richardson, T. Miller, G. Mayfield, P. Schafer, and J. Keung, "Yellowbrick."

[28] M. Claesen and B. D. Moor, "Hyperparameter Search in Machine Learning," 2015.

[29] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc 22Nd ACM SIGKDD Int Conf Knowl Discov Data Min*, ser. KDD '16. New York, NY, USA: ACM, 2016.

[30] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang, and Z. Zhao, "Detecting work-related stress with a wearable device," *Comput Ind*, vol. 90, 2017.