IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 4, 2021, accepted June 16, 2021, date of publication June 28, 2021, date of current version July 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093216

Cognitive Load Monitoring With Wearables–Lessons Learned From a Machine Learning Challenge

MARTIN GJORESKI[1,2], BHARGAVI MAHESH[3], TINE KOLENIK[1], JENS UWE-GARBAS[3], DOMINIK SEUSS[2], HRISTIJAN GJORESKI[4], MITJA LUŠTREK[1], MATJAŽ GAMS[1], (Member, IEEE), AND VELJKO PEJOVIĆ[5]

[1]Jožef Stefan Institute and Jožef Stefan Postgraduate School, 1000 Ljubljana, Slovenia
[2]Faculty of Computer Science, Università della Svizzera italiana, 6900 Lugano, Switzerland
[3]Intelligent Systems Group, Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany
[4]Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia
[5]Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia

Corresponding author: Martin Gjoreski (martin.gjoreski@ijs.si)

This work was supported in part by the Slovenian Research Agency (ARRS) under Grant U2-AG-16/0672, 0287, Grant P2-0098, and Grant N2-0136; in part by the Jožef Stefan Institute and Fraunhofer IIS through the German Academic Exchange Service and ARRS, in 2019, under Grant 57451027; and in part by the by the Wide Health project-Horizon 2020 under Grant 95227.

ABSTRACT To further extend the applicability of wearable sensors, methods for accurately extracting subtle psychological information from the sensor data are required. However, accessing subjective information in everyday life, such as cognitive load, remains challenging. To bring consensus on methods for cognitive load monitoring, a machine learning challenge is organized. The participants developed machine learning methods for cognitive load classification using wrist-worn physiological sensors' data, namely heart rate, R-R intervals, skin conductance, and skin temperature. The data from subjects solving cognitive tasks of varying difficulty is used for the challenge. This article presents a systematic comparison and multi-strategic performance evaluation of the thirteen methods submitted to this challenge. A systematic comparison of preprocessing techniques, classification algorithms, and implementation techniques is presented. Performance variations for different task difficulty levels, different subjects, and different experiment periods are evaluated. The results indicate that the most robust methods used multimodal sensor data, classical classification approaches such as decision trees and support vector machines or their ensembles, and Bayesian hyperparameter optimization for hyperparameter tuning. The most accurate models used handcrafted features that are further selected using sequential backward floating search and evaluated using stratified person-aware cross-validation strategy. Moreover, the results indicated better classification performance for specific test subjects, the tasks with the highest difficulty, and in some cases, the time elapsed since the start of the experiment. This dependency is likely due to model overfitting or due to the subjective nature of the psychophysiological process. The intersubject variability in responses is challenging to be captured through objective binary labels for cognitive load, thereby warranting more sophisticated annotation approaches.

INDEX TERMS Cognitive load, machine learning, wearable sensors.

I. INTRODUCTION

The availability of small, wearable, and low-cost sensors combined with advanced signal processing and information extraction capabilities is driving the revolution in mobile behavior monitoring for applications such as sports analytics, ambient-assisted living, and lifestyle monitoring [1]. The applicability of wearable sensors is enhanced by the extraction of subtle physiological information that can serve as the basis of psychological monitoring. However, assessing

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

psychophysiological information in everyday life remains challenging [2] since the association of wearable sensor data to human psychophysiological states is not as explicit as it is for physical states. For instance, smartphones can count steps and distinguish human physical activities (e.g., running vs. walking), but cannot recognize emotions and other affective states (e.g., cognitive load). Additionally, the inability of humans to recognize their own psychophysiological states in a timely and accurate manner poses a challenge for the development of affect recognition systems.

The psychophysiological state addressed in this paper is the cognitive load. It refers to the state of utilization of one's mental resources and is strongly related to attention. Mental resources are limited. A mentally-demanding task deprives the new tasks of resources. Consequently, the person cannot pay attention to these new tasks or must interrupt the current task. Wearable devices and mobile applications should be aware of the user's cognitive load when the user is occupied with a demanding task. This can prevent undesirable effects of attention-grabbing. For instance, nearly 25,000 lives are lost annually on the EU roads where a vast majority of accidents are caused by human error, often by a distracted driver.[1] Intelligent solutions to detect cognitive load and other mental states, and provide a warning when needed, may decrease the loss of human lives, thereby contributing to the EU's goal of zero fatalities and severe injuries by 2050.[2] Additionally, monitoring affective states can help improve mental well-being [4] and productivity (e.g., avoiding notifications while the user is in the optimal flow state) [5].

When humans experience a psychophysiological load in the form of a demanding task, the sympathetic nervous system is activated. Depending on the load intensity, this activation increases the heart rate, sweating rate, breathing rate, and blood pressure; the pupils dilate, the saliva flow decreases, the heartbeats become equidistant, the blood flow is restricted from the extremities, and is redirected towards the vital organs. These signals can be measured accurately in controlled environments, such as hospitals, using specialized equipment. However, less obtrusive and less expensive devices are required to capture these signals in daily life through practical and large-scale experimentation [7]. Moreover, an ecological momentary assessment that reveals user experiences are necessary to infer mental states from such measurements in daily life [6]. Recent advances in sensing technology have enabled relatively unobtrusive vital sign monitoring, thereby, bringing us closer towards unobtrusive mental state monitoring [26]. A significant part of research in mental state recognition and monitoring with wearables focuses on mental stress. For instance, Mozos *et al.* [54] used wearable and sociometric sensors to detect stress using a standard stress induction protocol. Similarly, Gjoreski *et al.* [30] used commercially available Empatica wristbands to detect stress with up to 92% accuracy using heart rate variability, blood volume pulse, galvanic skin response (GSR), skin temperature, and acceleration. Stress often overlaps with the cognitive load but can be potentially distinguished from it [27]. Inferencing cognitive load from physiological signals is an important research field that is less researched compared to the recognition of physical states and activities, as well as the inference of several psychological states (e.g., stress, affect). To promote this field, a machine learning (ML) challenge was organized in which the participants built pipelines to infer cognitive load. Since the same dataset was used for the ML pipelines, performances of the algorithms could be compared and the best methods for cognitive load inferencing could be ascertained.

This article has the following contributions: i) it presents a systematic comparison of approaches of the thirteen successful machine learning pipelines submitted to the aforementioned challenge, ii) it provides a detailed evaluation of their overall performance and their performances for different subjects, different tasks and their difficulty levels, and iii) it summarizes the learnings from the challenge and presents them as suggestions for ML model development to infer cognitive load.

FIGURE 1. Wristband microsoft band 2 used for dataset collection.

II. CHALLENGE DATASET DESCRIPTION

In order to collect physiological signals in situations where a subject is cognitively engaged, an experiment was conducted in which the subjects solved cognitive tasks of varying difficulty. The experiment was performed in a quiet, normal-temperature office with one subject at a time under the same circumstances. Twenty-three subjects (four female) were recruited through the institutional communication channels (e.g. mailing lists, social network posts) and personal links. Their mean age was 29.5. The subjects had various degrees of educational qualification – high school (7), B.Sc. (6), M.Sc. (6), and Ph.D. (4). studies. All subjects were (self-assessed) healthy adults and no other criteria were used for limiting the participation. The subjects wore a commercial wristband (refer Figure 1) on their non-dominant arm and sat on a comfortable chair in front of a computer monitor. The experiment session was recorded without any restrictions on the subject's hand gestures, thereby reproducing sedentary workstyle. The experiment protocol is depicted in Figure 2. The subjects were briefed about the experiment. The remaining protocol comprised of two sets of tests – cognitive capac-

[1] https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1793

[2] https://www.ubittention.org/2020/data/Cognitive-load challenge description.pdf

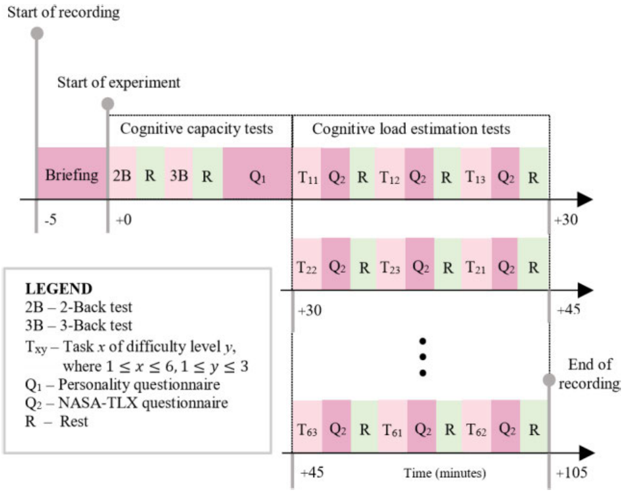

Start of recording
Start of experiment
Cognitive capacity tests
Cognitive load estimation tests
Briefing
2B
R
3B
R
Q1
T11 Q2 R T12 Q2 R T13 Q2 R
-5
+0
+30
T22 Q2 R T23 Q2 R T21 Q2 R
+30
+45
End of recording
T63 Q2 R T61 Q2 R T62 Q2 R
+45
Time (minutes)
+105
LEGEND
2B – 2-Back test
3B – 3-Back test
Txy – Task x of difficulty level y, where 1 ≤ x ≤ 6, 1 ≤ y ≤ 3
Q1 – Personality questionnaire
Q2 – NASA-TLX questionnaire
R – Rest

FIGURE 2. Dataset collection protocol.

ity tests and cognitive load estimation tests. A demographic questionnaire was filled in between the two tests. Cognitive capacity tests consisted of n-back tasks where n ∈ {2, 3} (2B and 3B in Figure 2). An n-back task consisted of 3 × 3 grid cells, one of which was colored at each time step. The subjects decided whether the colored cell at a time step was the same as the one colored *n* steps ago. The ratio of correct and incorrect answers depicted the cognitive capacity of the subject. Cognitive load estimation tests were comprised of six elementary cognitive tasks (ECTs) (denoted by x in T_{xy} in Figure 2). These tasks are designed to elicit perceptual cognitive engagement, often used to demonstrate individual differences among people [35]. Haapalainen *et al.* [9] developed a software with these ECTs to assess visual-perception-based cognitive load factors. A variation of this software was utilized for the data collection. The six ECTs were: i) Gestalt Completion test (T_1) to identify incomplete drawings, ii) Hidden Pattern test (T_2) to identify if a given model image is hidden in the composition of other images, iii) Finding A's test (T_3) to capture the speed of identification of letter 'a's in a text, iv) Number Comparison test (T_4) to gauge the subject's speed of comparison of two multidigit numbers, v) Pursuit test (T_5) to visually track irregularly-curved overlapping lines from the numbers on left to letters on the right side of a rectangle, and vi) Scattered X's test (T_6) to find the letter 'x' placed randomly, crowded with other letters. The first four ECTs were obtained from a manual for reference tests for cognitive factors [36], a popular standard for educational psychology research. The last two ECTs were originally devised by Thurstone and Thurstone [37]. Furthermore, each ECT had three variations in difficulty (easy, medium and hard difficulty levels denoted by y in T_{xy} in Figure 2) and were presented in a randomized order. After each task, a NASA-TLX [8] questionnaire was filled by the subjects to assess subjective cognitive load. The participants rested for three minutes after filling each questionnaire.

The following wristband data was recorded with 1Hz sampling rate: R-R (or inter-beat) intervals, galvanic skin response (GSR), heart rate (HR), skin temperature (ST), barometer data, accelerometer and UV index data. However, the focus of the challenge is limited to the data from the following physiological sensors: R-R, GSR, ST, and HR. The data from the wristband was transmitted via Bluetooth and a mobile phone to a server for offline data analysis. Figure 3 depicts the signals for a subject in a single session.

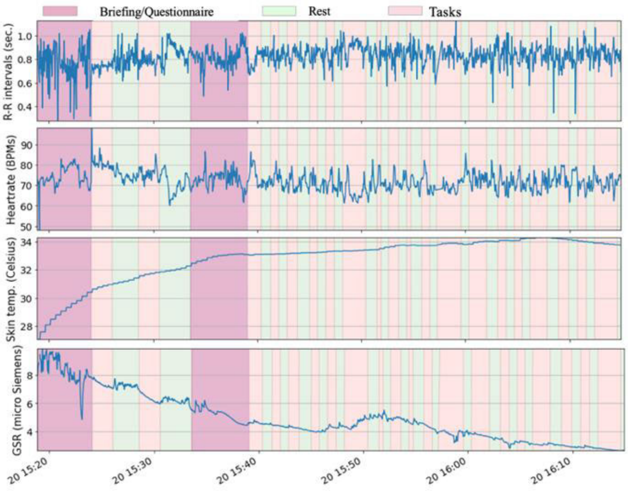


FIGURE 3. Sample sensor data for a subject. NASA-TLX questionnaire periods have been excluded.

Due to excessive noise, affected segments in the original dataset were disregarded. The dataset used for the challenge consisted of 825 instances from 23 participants. The instances of rest were labelled as 'no load' whereas the task instances were labelled as 'cognitive load'. Each instance was composed of 30-seconds data of four modalities: R-R, GSR, ST, and heart rate. The dataset was split into training and test datasets with 632 instances from 18 subjects in the former. In the training set, 49.6% of the instances had a label '0' or 'no load, hence leading to a nearly balanced dataset. Each subject's data was assigned a unique subject ID. Furthermore, the dataset is the first labeled dataset for cognitive load monitoring with a wristband and is made publicly available following the ML challenge.

III. MACHINE LEARNING CHALLENGE

The goal of this challenge was to recognize two levels of cognitive load – Cognitive load vs. no load, using four physiological signals – R-R, GSR, ST, and HR. The participants of the challenge had access to a labeled training dataset and an unlabeled test dataset. The participants developed ML pipelines that processed the sensor data, created models, and recognized the cognitive load. The problem is deliberately reduced to the binary recognition of whether a subject is engaged in a task (irrespective of whether the task is easy, medium, or hard) or resting, as the previous efforts demonstrate that fine-grain distinction among different cognitive load levels

TABLE 1. Comparison of the methods adopted by the challenge participants.

Method	Preprocessing	Features	Feature Selection	Proposed Classifier
I	-	Handcrafted: general	Sequential backward floating search	Ensemble of 7 Gradient boosting decision trees
II	-	Handcrafted: features in [34]	-	Support vector machine
III	Standardization (subjectwise)	Handcrafted: general and domain-specific	Sequential forward floating search	Ensemble of support vector machines [33]
IV	Standardization	Handcrafted: general and domain-specific	-	Logistic regression
V	Min-max normalization (overall and subjectwise)	Handcrafted: general and domain-specific (partially automated)	Feature discovery platform	Random forest
VI	Standardization	Handcrafted: general and domain-specific	-	Weighted sum of individualized and global logistic regression models
VII	Min-max normalization	Handcrafted: general and domain-specific	Maximal information coefficient	Multilayer perceptron
VIII	-	Handcrafted: general and domain-specific	-	Logistic regression
IX	Standardization (subjectwise)	Handcrafted: general and domain-specific	Gini impurity	Support vector machine [37]
X	-	Handcrafted: correlation dimension	-	XGBoost classifier
XI	Standardization	Automated: pretrained on a larger dataset [35]	-	Convolutional neural network (6 layers per sensor data, each layer with batch normalization and ReLU activation)
XII	Min-max normalization (subjectwise)	Automated	-	Recurrent neural network [36]
XIII	Standardization (subjectwise)	Handcrafted: general and domain-specific	Gini impurity	Logistic regression

from physiological signals might be impossible [26], [32]. The results were presented at UbiTtention workshop at ACM UbiComp 2020 conference, and the three best-performing teams were rewarded. The following subsections describe the specifications of ML pipelines submitted to this challenge in further detail.

A. METHODS

This subsection describes the methods adopted by the participants of the ML challenge to infer cognitive load. The challenge received thirteen submissions from nine different teams. In the following sections, each submission is regarded as a method and denoted by a roman number. Further details on the teams are provided in the Appendix. Table 1 provides an overview of the methods. Nine methods involved preprocessing techniques such as standardization or normalization. Notably, more than half of them used subjectwise preprocessing. A majority, i.e., ten out of thirteen methods, are based on classical ML approaches, including tree-based algorithms (I, V, VI, X), support vector machines and their ensembles (II, III, IX), and logistic regression (IV, VIII, XIII). The remaining three are based on neural networks: a multilayer perceptron (VII), a recurrent neural network (XII), and an autoencoder based on a convolutional neural network (XI). However, only two of these three are end-to-end learning approaches. The small dataset size was noted as a major motivation for choosing classical ML approaches over approaches based on neural networks. To overcome the shortcoming posed by the dataset size during training, three methods adopted dataset augmentation techniques, whereas the transfer-learning-based approach in method XI used an external, yet similar dataset to pretrain the model. Method VII utilized Synthetic Minority Over-sampling Technique (SMOTE) to enlarge the dataset as well as to introduce variability. Meanwhile in method XII, a particular class was upsampled to counteract the input-induced bias in the network. Method X used B-spline interpolation of instances to compensate for the effects of low sampling frequency. All the methods considered the four modalities provided.

A majority (eleven) of the methods involved handcrafted feature extraction. Among the extracted features, the prominent ones encompassed time-domain statistical measures such as mean, variance, kurtosis, median, sum, etc. and frequency-domain measures such as power spectral density ratio of heart rate variability. Several extracted features are modality-specific, e.g., skin conductance peak amplitudes are derived from GSR, and heart rate variability in terms of root mean square of successive differences derived from

R-R intervals. The total number of extracted features varied between 4 and 129. However, six approaches did not utilize all the extracted features. Instead, feature selection techniques such as maximal information coefficient, sequential forward or backward floating selection, and Gini impurity method are used to select the most informative features. Method V performed feature extraction and selection using an in-house feature discovery platform. ML algorithms rely heavily on hyperparameters. Hence, hyperparameter optimization plays a vital role. Three methods optimized the hyperparameters with a grid search (IX), Bayesian optimization (I), and their combination (III).

B. IMPLEMENTATION FRAMEWORK

Python is the most prominent programming language used by the participants and the *scikit-learn* library is commonly used for classical ML algorithms. The hyperparameter optimization library *hyperopt* is utilized in two methods. The models are internally evaluated on a validation set. Twelve out of thirteen methods have mentioned the use of a cross-validation strategy for evaluation. Most of them used the leave-k-subjects-out strategy, while others used a leave-k-folds-out strategy or a combination of both (refer Table 2). The resulting models vary in size depending on the algorithm. The logistic regression model developed in method IV resulted in the smallest size (845 B), whereas the convolutional neural network model developed in method XI resulted in the largest size (37 MB).

TABLE 2. Method implementation tools and evaluation strategies based on information provided by participants.

Method	Tools/framework	Evaluation strategy	Model Size (bytes)
I	scikit-learn, pandas, lightgbm, hyperopt.	5-Fold CV	3 M
II	mlxtend	Leave-1-subject-out	-
III	tsfresh, scikit-learn, mlxtend, hyperopt	Leave-3-subjects-out Leave-1-subject-out	1 M
IV	scikit-learn, PyWavelets, SciPy, eda-explorer	5-fold CV	845
V	MATLAB, Signal Properties based Generic Features.	10-fold CV Leave-2-subjects-out	5 M
VI	scikit-learn, PyWavelets, SciPy, eda-explorer	Leave-5-subjects-out	7 K
VII	scikit-learn, pandas, SciPy	Leave-4-subjects-out	19 K
VIII	scikit-learn	10-fold nested CV	-
IX	scikit-learn, SciPy	6-fold CV Leave-3-subjects-out	279 K
X	-	-	-
XI	Tensorflow, mlxtend	Leave-1-subject-out	37 M
XII	Keras	Leave-3-subjects-out	34 K
XIII	scipy, scikit-learn	6-fold CV Leave-3-subjects-out	4 K

IV. CLASSIFICATION PERFORMANCE EVALUATION

We evaluated the methods on the test dataset using various strategies:

i) *Overall Classification Performance:* Average binary classification accuracies of the methods on the entire test dataset are computed. Further, the highest achievable performance is obtained through voting ensembles of multiple methods.

ii) *Subject-Related Performance:* Binary classification accuracy is computed for the five test subjects. This evaluation strategy potentially depicts the user-generalization capability of the model.

iii) *Task-Difficulty-Related Performance:* This strategy focuses on binary classification accuracy for the three task difficulty levels. This strategy depicts the variation of classification complexity based on task difficulty.

iv) *Experiment-Period-Related Performance:* This strategy focuses on binary classification accuracy for each of the two halves of the experiment period, potentially depicting the influence of the duration of the experiment on the performance of the model.

A. EVALUATION METRICS

The methods are evaluated on the instances in the test dataset. One of the following two performance metrics is used depending on the aforementioned strategies: accuracy (Acc) for the first evaluation strategy and partial accuracy (pAcc) for the remaining strategies. Accuracy is the standard ML score defined as:

$$Acc = \frac{\#\, correctly\ predicted\ instances}{\#\, instances}$$

Partial accuracy is used for the remaining evaluation strategies, and is accuracy calculated over a subset of instances x as:

$$pAcc(x) = \frac{\#\, correct\ pred.\ for\ the\ instances\ from\ x}{\#\, instances\ from\ x}$$

Depending on the evaluation strategy, x can represent any of the following: instances from a test subject, instances from a task with specific difficulty (e.g., rest, easy, medium or hard), or instances from a portion of the experimental period (e.g., first half vs. second half). Though the ML methods are initially developed for binary classification (rest vs. cognitive load), the partial accuracy allows for a better granularity in the analysis of the methods. Additional evaluation scores such as precision, recall, and F1-score for overall performance are presented in the appendix.

B. INFERENCE ACCURACY

Table 3 presents the average accuracy achieved by each method on the test dataset. The accuracies spread gradually from baseline 0.5 to the highest accuracy of 0.69. However,

TABLE 3. Overall classification performance: average accuracy of methods on the test data in decreasing order.

Method/Rank	Accuracy
I	0.694
II	0.679
III	0.674
IV	0.663
V	0.653
VI	0.653
VII	0.648
VIII	0.648
IX	0.627
X	0.580
XI	0.560
XII	0.554
XIII	0.503

none of the methods significantly outperformed the remaining. The top-ranked method resulted in an accuracy of 0.694, which is 0.15 higher than the second-best method and 0.2 higher than the third-ranked method.

Table 4 presents the accuracies achieved by voting ensembles of the top-x ranked methods. The highest accuracy of 0.71 is achieved using a voting ensemble of the top-3 methods.

TABLE 4. Accuracy achieved by voting ensembles of top-x methods. votes from top 3 methods result in the best performance.

	Voting of Top-x						
x	1	3	5	7	9	11	13
Accuracy	0.694	0.710	0.694	0.668	0.679	0.674	0.679

Table 5 presents the partial accuracy per subject in the test dataset for each of the methods. The results are seen to be subject-dependent and most of the methods perform well for specific subjects (e.g., subjects with IDs *iz3x1* and *bd47a*). For subjects *3caqi* and *f1gjp*, most of the methods do not perform well. The dependency on the subjects is less obvious for the higher-ranked methods than for the lower-ranked methods. For instance, method I achieved the highest accuracy of 0.789 and the lowest accuracy of 0.615, resulting in a difference of 0.174. This difference is much higher for the rest of the methods, including the second-ranked and the third-ranked methods. This indicates good user-generalization capabilities of method I.

TABLE 5. Subject-related performance: partial accuracy of each method per test subject.

	Subject ID				
Method	*bd47a*	*6frz4*	*iz3x1*	*3caqi*	*f1gjp*
I	0.711	0.718	0.789	0.641	0.615
II	0.763	0.744	0.763	0.641	0.487
III	0.816	0.692	0.658	0.564	0.641
IV	0.763	0.641	0.895	0.564	0.462
V	0.842	0.667	0.763	0.513	0.487
VI	0.737	0.590	0.895	0.641	0.410
VII	0.816	0.615	0.737	0.615	0.462
VIII	0.684	0.692	0.737	0.564	0.564
IX	0.789	0.564	0.632	0.538	0.615
X	0.658	0.615	0.605	0.564	0.462
XI	0.632	0.615	0.711	0.359	0.487
XII	0.447	0.718	0.526	0.564	0.513
XIII	0.421	0.538	0.421	0.410	0.718

TABLE 6. Task-difficulty-related performance: partial accuracies per task difficulty. Better classification performance is observed for harder tasks.

	Designed Task Difficulty			
Method	Rest	Easy	Medium	Hard
I	0.632	0.679	0.714	0.857
II	0.600	0.714	0.771	0.771
III	0.695	0.679	0.686	0.600
IV	0.632	0.607	0.657	0.800
V	0.663	0.571	0.657	0.686
VI	0.579	0.571	0.686	0.886
VII	0.611	0.643	0.600	0.800
VIII	0.642	0.714	0.629	0.629
IX	0.621	0.571	0.714	0.600
X	0.674	0.429	0.457	0.571
XI	0.453	0.607	0.657	0.714
XII	0.632	0.643	0.457	0.371
XIII	0.516	0.607	0.343	0.543

Table 6 presents the partial accuracy per designed task difficulty. The results show that most of the high-ranked methods perform better for the instances belonging to higher task difficulty. The exceptions to this are methods III, V, and XI. The rest periods are the most challenging to detect for all of the methods. Since the difficulty levels are presented in a random order, the rest periods are further analyzed by segregating them based on the preceding task difficulty to identify whether the prior difficulty influences the accuracy of rest detection.

Table 7 presents the partial accuracy for rest periods followed by easy, medium, and hard tasks. It can be seen that there is no specific pattern depicting the influence of task difficulty on rest period accuracies.

TABLE 7. Partial accuracies for rest periods following different task difficulties.

	Rest Period		
Method	Rest after Easy task	Rest after med. task	Rest after hard task
I	0.593	0.667	0.600
II	0.519	0.636	0.600
III	0.741	0.727	0.667
IV	0.593	0.636	0.633
V	0.778	0.576	0.600
VI	0.519	0.636	0.533
VII	0.630	0.636	0.533
VIII	0.556	0.667	0.667
IX	0.481	0.636	0.667
X	0.667	0.697	0.633
XI	0.333	0.576	0.367
XII	0.741	0.667	0.467
XIII	0.444	0.636	0.467

TABLE 8. Experiment-period-related performance: partial accuracies for experiment period halves.

	Experiment Period	
Method	First half	Second half
I	0.682	0.716
II	0.671	0.682
III	0.706	0.636
IV	0.694	0.625
V	0.671	0.659
VI	0.671	0.602
VII	0.647	0.659
VIII	0.635	0.659
IX	0.647	0.625
X	0.624	0.523
XI	0.588	0.489
XII	0.541	0.534
XIII	0.518	0.500

Table 8 presents the partial accuracy with respect to the experiment period, i.e., the first half of the experiment vs. the second half of the experiment. The results show that methods such as III and IV are sensitive to the experiment period as they have larger variation in the accuracies achieved for the two halves of the experiment in comparison with the other methods.

C. POSSIBLE CAUSES OF OVERFITTING

Multi-strategic evaluation of models uncovered possible influences of training/test splitting on the performance. Table 6 depicted the dependency of methods' performance on the subjects in the test dataset. The inter-subject performance variation is lower for the top-ranked methods, indicating higher generalizability. Performance variation of lower-ranked methods is likely a sign of overfitting, which needs to be considered by the researchers during model selection. One possible solution is to include the inter-subject performance variation as an additional optimization parameter during the model training. Results in Table 8 depicted higher sensitivity of low-ranked methods to the experimental period. This additionally indicates overfitting where the experiment design influenced the ML models. Possible solutions to these problems include optimal tuning of the ML models and better feature selection methods to remove the features sensitive to the experiment period.

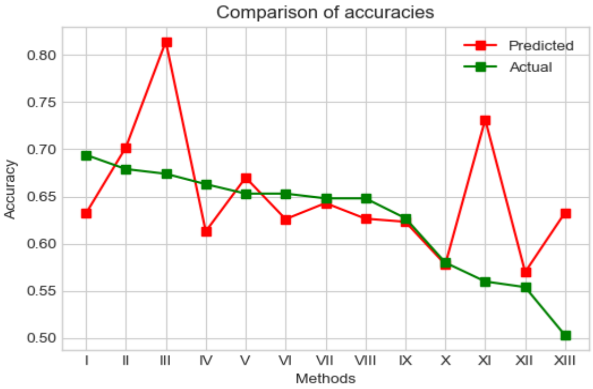


FIGURE 4. Overview of validation (predicted) accuracies of submitted models and the corresponding accuracy on the test set (actual).

Finally, a higher predicted accuracy achieved on a validation set (or using the cross-validation on the train data) compared to the test set accuracy indicates overfitting (refer Figure 4). Such overfitting may appear when hyperparameter tuning is performed using the same cross-validation scheme that has been used for evaluating the final models. A possible solution to this problem for small datasets could be a nested cross-validation approach. For larger datasets, the traditional train-validation-test splits are often sufficient.

D. METHOD SIMILARITY

Performing a statistical-significance analysis over the presented results is challenging since the methods are tested only once on the final test data. To present some intuition about the differences in methods, we performed hierarchical clustering using Euclidean distance and complete linkage applied over the methods' predictions (refer Figure 5).

In Figure 5, the end-to-end learning methods (XI and XII) are partitioned out of homogenous clusters, depicting that they are not identical to the feature-engineering-based methods. Additionally, the four of the top-5 (I, II, III, and V) belong to a same cluster.

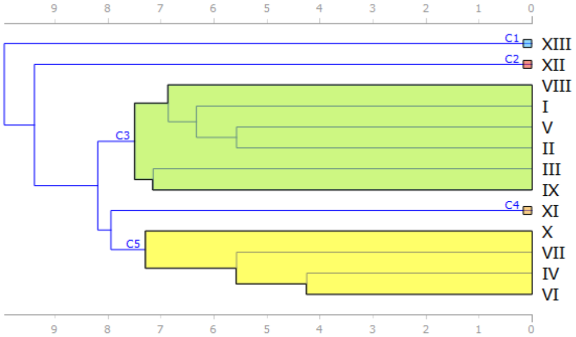


FIGURE 5. Hierarchical clustering using Euclidean distance and complete linkage applied over the methods' predictions.

V. DISCUSSION AND LESSONS LEARNED

Our meta-analysis presented in the previous section reveals the superiority of a combination of data processing techniques for the wrist-worn device-originated physiological signals for cognitive load inference. Namely, we observe that ensemble-based ML algorithms in conjunction with sequential backward floating search feature selection, Bayesian hyperparameter optimization, and evaluation founded in stratified person-aware cross-validation outperform alternative approaches.

To move beyond the competition of different methods, and to guide future efforts in automated cognitive load inference, certain peculiarities of sensor data elicited during human cognitive engagement are listed below. They imply a particular manner in which cognitive load inference pipelines should be constructed. The inferences are as follows: i) Physiological response to increased cognitive load is relatively subtle, represented by changes that may be symptomatic to other phenomena (e.g. a subject's health status, emotions, physical stress, etc.), and prone to noise, especially when collected via inexpensive wearable sensors. Consequently, while deep-learning-based automatic feature extraction excels in several other domains, cognitive load inference still requires carefully handcrafted features and guided feature selection to avoid the algorithm's attention on irrelevant signals. Naturally, the three neural network-based submissions are among the low-ranked methods. ii) The methods analyzed in this paper perform relatively well when a subject is highly cognitively engaged yet fail when the subject is resting or engaged in an easy task. It appears that the physiological signal variation captured by commercial wearable devices is rather minuscule to allow fine-grain detection of cognitive load levels. These findings are in line with the related work [30], [32]. iii) Subject-related analysis reveals that one solution that fits all may not be feasible. Different approaches are successful when inferring the cognitive engagement of different subjects. Confounding variables likely related to a subject's demographics or personality may result in different physiological reactions. Hence, the development of a suitable ML model for a particular subject is an interesting avenue for future research. iv) The analyses demonstrate the need for a separate well-founded evaluation set when physiological signals are considered. Despite the popularity and practicality of cross-validation, independent evaluation with well-stratified data initially separated from the training set is crucial to avoid unintentional overfitting.

Besides the observations presented so far, it should be noted that additional challenges exist for an in-the-wild cognitive load monitoring system. The dataset analyzed in this study was collected in a sedentary environment. On the other hand, Schmalfus *et al.* [14] explored the potential of wearable devices for mental workload detection in different physiological activity conditions. The study included 32 participants, 2 mental stressors and 4 physical stressors. The statistical analysis indicated that wearable devices are not fully capable of identifying mental workload when physical activity is present.

The tasks of our data collection experiments are geared specifically towards eliciting different levels of cognitive load. These tasks have been a part of the standard psychological toolbox since the 1940s VII, and their implementation (introduced by Haapalainen *et al.* [9]) used in this work has been considered by other studies as well (e.g., [26] and [16]), affirming that the stimulus of the experiment protocol was indeed cognitive load.

Physiological signals captured by the Microsoft Band wristband include heart activity-related signals, acceleration, skin temperature, and skin conductance. More than one confounding factor may affect the change in these signals. For instance, heart activity can increase due to a subject's health state, emotion, stress, and other factors. However, the relationship between the heart activity-related signals and cognitive load is well-documented in the existing literature (e.g. [16], [23]). To a certain extent, the relationships between cognitive load and skin conductance (e.g., [17] and [18]), as well as the skin temperature [16]) have also been researched.

VI. RELATED WORK

A variety of psychophysiological measures can be used for assessing cognitive states: electroencephalography (EEG), electrocardiogram (ECG), heart rate and heart rate variability, optical imaging, blood pressure, skin conductance, electromyography, thermal imaging, pupilometry [10]. The majority of the efforts related to cognitive load monitoring with wearable sensors, however, focused on EEG devices. This is a natural choice as the brain is the most informative source of information for monitoring human psychological states using sensors. Usually, features are extracted from the EEG sensor data (e.g., the intensity of different frequency bands), and those features are analyzed using correlation analysis [11] or ML models (Naive Bayes, Linear Discriminant Analysis, SVM, Convolutional Neural Network – CNN, Logistic Regression) [19], [20], [25]. Moving further towards multimodal sensing, Jimenez-Molina *et al.* [23] explored photoplethysmography (PPG), EEG, temperature and pupil dilation sensors to the assess mental workload of 61 participants during web browsing. Contrary to the studies based on physiological sensors, Chen and Epps [21] used gyroscope-based atomic head movement analysis for task load recognition. All of these studies involving EEG and head mounted-devices can be quite useful for cognitive load monitoring in movement-restricted environment, such as in virtual-reality-based scenarios, but their application remains limited in real life.

Additionally, chest-mounted devices, which are less obtrusive than head-mounted devices, have been proven useful for cognitive load monitoring [9], [13], [15], [22] though accompanied by real-life limitations.

Compared to head- and chest-mounted devices, wrist-worn devices are likely the least obtrusive because subjects are already accustomed to wearing wristwatches. Johannessen *et al.* [13] analyzed cognitive load in 5 physician team leaders during trauma resuscitation. They collected glasses-based eye-tracking data and wrist-based GSR, and heart rate data, during five trauma resuscitations. A correlation and regression analysis showed that multiple physiological measures should be employed to most accurately measure cognitive load in a real-world setting. Kohout *et al.* [24] proposed an approach for detecting cognitive load (relaxed vs. loaded) by collecting data from 8 participants wearing wrist sensors and additionally carrying a smartphone as a sensor in their pocket while performing a pill-sorting task. They stressed their participants by introducing a dual-task situation. They used an SVM classifier to achieve 90% accuracy. Novak *et al.* used wristbands to infer cognitive load in a simulated driving environment [28]. Similarly, Gjoreski *et al.* combined physiological sensors with video-based sensors to detect increased cognitive load while driving [31]. Schaule *et al.* [29] used the same wristbands and an N-back task to elicit different levels of cognitive load among office workers.

Barua *et al.* [38] used the n-back task to assess cognitive load in drivers while measuring their physiological signals (ECG, GSR, respiration, EEG, electrooculography). The authors used various ML models, including k-nearest neighbor (k-NN), SVM, and random forest for classifying cognitive load, and random forest outperformed other methods. Yomna *et al.* [40] collected measurements on eye movements in drivers and compounded them with data on braking, acceleration and steering. Reasonable accuracies were obtained by using SVM and random forest methods for recognizing abnormal driving situations through the cognitive load of drivers. Fridman *et al.* [41] tried to estimate cognitive load in real-life driving situations by employing vision-based methods, captured in a video. The best-implemented method with high accuracy was a 3D convolutional neural network. Appel *et al.* [42] experimented with participants in various game simulation environments, collecting data on interaction metrics, pupil dilation, eye-fixation behavior, and heart rate data. Participant-specific random forest achieved the best accuracy in classifying cognitive load. Chen *et al.* [43] measured cognitive load by four methods: the subjective rating of task difficulty, task completion time, performance accuracy and eye activity-based physiological measurement. ANOVA tests and Gaussian mixture model classification resulted in the best classification accuracy in classifying five levels of cognitive load. The authors noted that eye activity is the best measure for cognitive load due to real-time accessibility. Nourbakhsh *et al.* [44] focused on GSR and eye blinks as their measurements for the cognitive load. The participants in the study took an arithmetic test with four different difficulty levels while the measurements were taken. Naive Bayes achieved the best accuracy for binary classification, while SVM achieved the best accuracy for 4-level classification. Yin *et al.* [45] estimated three different levels of cognitive load from speech in a speaker-independent setting. The best accuracy was produced by a Gaussian mixture model with 256 mixtures using a background model with maximum a-posteriori estimation technique for different levels of cognitive load, using Mel-Frequency Cepstral Coefficients, prosodic features, acceleration features, and feature warping. Segbroeck [46] extracted static and dynamic features from speech to estimate three levels of cognitive load. By performing a feature-level fusion on various features (prosodic, spectral, voice quality, lexical information, speaking rate) with i-vector modelling, they produced better results than existing SVM models.

Furthermore, the least obtrusive approaches are those approaches that infer cognitive load using remote sensing [26], [40] although they are challenging. Cognitive load inference may also be beneficial in the future for people with various brain-related disorders, e.g. Parkinson's disease or multiple sclerosis [51], [52].

All of these studies demonstrate the usability of wearable sensors for monitoring cognitive load and related psychophysiological constructs (e.g., stress, distractions, etc.). Typically, in all of these studies, one novel approach is compared against a few baselines on a dataset that is not publicly available. In our study, thirteen novel methods were analyzed and evaluated against the same benchmark data, which is publicly available, thus allowing for reproducible and systematic advancement of the field.

VII. CONCLUSION

In this paper, we analyzed thirteen methods for cognitive load inference from wrist-worn physiological sensors that were submitted to an online ML challenge. The methods were compared and evaluated against the same benchmark data, and a systematic comparison was presented with respect to preprocessing techniques, dataset augmentation techniques, extracted features, feature selection algorithms, classification algorithms, hyperparameter optimization techniques, evaluation approaches, and technical implementation. This work also evaluated the impact of different task difficulty levels, different subjects, and different experiment periods on classification performance. Based on this performance evaluation, the most promising data processing blocks, including

TABLE 9. Summary of the findings and guideline for accurate cognitive load monitoring models.

Dataset augmentation	No significant role
Preprocessing	Subjectwise standardization
ML approach	Ensemble of ML models
Features	Time-frequency domain, statistics, HRV
Feature selection	Sequential backward floating search
Hyperparameters	Bayesian optimization
Evaluation	Stratified subject-aware cross-validation
Other	- High-ranked methods have lower inter-subject accuracy difference. - High-ranked methods perform better for the instances that have a higher designed task difficulty. - Low-ranked methods are more sensitive to the different experiment periods.

TABLE 10. Teams participating in cognitive load monitoring challenge and their rankings.

Method	Ranking	Team Name	Affiliation
I	1	major_tom	University of Tuebingen
II	2	HCM-feature	Augsburg University
III	3	Smart D.-D. L	VTT Technical Research Centre
IV	4	IdeasLabUT_1	University of Toledo
V	5	TCS	TCS Research & Innovation
VI	6	IdeasLabUT_2	University of Toledo
VII	7	TCS 2	TCS Research & Innovation
VIII	8	Janus	Indian Institute of Technology Delhi
IX	9	Lynix 1	University of Washington
X	10	DataVaders	Indian Institute of Technology Mandi
XI	11	HCM-auto	Augsburg University
XII	12	Sala	Johannes Kepler University Linz
XIII	13	Lynix 2	University of Washington

TABLE 11. Additional evaluation scores (accuracy, precision, recall and F1-score).

Method	Accuracy	Precision	Recall	F1 (micro)
I	0.694	0.714	0.632	0.694
II	0.679	0.704	0.600	0.679
III	0.674	0.660	0.695	0.674
IV	0.663	0.667	0.632	0.663
V	0.653	0.643	0.663	0.653
VI	0.653	0.671	0.579	0.653
VII	0.648	0.652	0.611	0.648
VIII	0.648	0.642	0.642	0.648
IX	0.627	0.621	0.621	0.627
X	0.580	0.561	0.674	0.580
XI	0.560	0.566	0.453	0.560
XII	0.554	0.541	0.632	0.554
XIII	0.503	0.495	0.516	0.503

classification algorithms, were identified and summarized in Table 9. Weiser's vision of a computer fully understandable of its subjects might appear to be wishful thinking in the early twenty-first century [48]. However, we believe that the identification of the most promising approaches for cognitive load inference that are demonstrated in this paper through an unbiased analysis of solutions submitted to a global machine learning challenge provides a sound basis for the future work towards the realization of this vision.

APPENDIX

See Tables 10 and 11.

ACKNOWLEDGMENT

(Martin Gjoreski and Bhargavi Mahesh contributed equally to this work.)

REFERENCES

[1] M. Cardinale and M. C. Varley, "Wearable training-monitoring technology: Applications, challenges, and opportunities," *Int. J. Sports Physiol. Perform.*, vol. 12, no. s2, pp. S2-55–S2-62, Apr. 2017.

[2] D. Heaven, "Why faces don't always tell the truth about feelings," *Nature*, vol. 578, no. 7796, pp. 502–504, Feb. 2020.

[3] B. Felix, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1–4.

[4] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervas. Mobile Comput.*, vol. 51, pp. 1–26, Dec. 2018.

[5] J. L. King and A. J. Ehrenberg, "The productivity vampires," *Inf. Syst. Frontiers*, vol. 22, no. 1, pp. 11–15, Feb. 2020.

[6] Y. Rogers, K. Connelly, L. Tedesco, W. Hazlewood, A. Kurtz, R. E. Hall, J. Hursey, and T. Toscos, "Why it's worth the hassle: The value of *in-situ* studies when designing Ubicomp," in *Proc. Int. Conf. Ubiquitous Comput.*, Berlin, Germany: Springer, 2007, pp. 336–353.

[7] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, and V. Pejović, "Datasets for cognitive load inference using wearable sensors and psychological traits," *Appl. Sci.*, vol. 10, no. 11, p. 3843, May 2020.

[8] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol*, vol. 52, pp. 139–183, Jan. 1988.

[9] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proc. 12th ACM Int. Conf. Ubiquitous Comput.*, Sep. 2010, pp. 301–310.

[10] M. Lohani, B. R. Payne, and D. L. Strayer, "A review of psychophysiological measures to assess cognitive states in real-world driving," *Frontiers Hum. Neurosci.*, vol. 13, p. 57, Mar. 2019.

[11] Y. Wu, T. Miwa, and M. Uchida, "Using physiological signals to measure operator's mental workload in shipping—An engine room simulator study," *J. Mar. Eng. Technol.*, vol. 16, no. 2, pp. 61–69, May 2017.

[12] K. Mohanavelu, S. Poonguzhali, D. Ravi, P. K. Singh, M. Mahajabin, K. Ramachandran, U. K. Singh, and S. Jayaraman, "Cognitive workload analysis of fighter aircraft pilots in flight simulator environment," *Defence Sci. J.*, vol. 70, no. 2, pp. 131–139, Mar. 2020.

[13] E. Johannessen, A. Szulewski, N. Radulovic, M. White, H. Braund, D. Howes, D. Rodenburg, and C. Davies, "Psychophysiologic measures of cognitive load in physician team leaders during trauma resuscitation," *Comput. Hum. Behav.*, vol. 111, Oct. 2020, Art. no. 106393.

[14] F. Schmalfuß, S. Mach, K. Klüber, B. Habelt, M. Beggiato, A. Körner, and J. F. Krems, "Potential of wearable devices for mental workload detection in different physiological activity conditions," in *Proc. Hum. Factors Ergonom. Soc. Europe*, 2018, pp. 179–191.

[15] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *Proc. 36th Int. Conf. Softw. Eng.*, May 2014, pp. 402–413.

[16] M. Myrtek, E. Deutschmann-Janicke, H. Strohmaier, W. Zimmermann, S. Lawerenz, G. Brügner, and W. Müller, "Physical, mental, emotional, and subjective workload components in train drivers," *Ergonomics*, vol. 37, no. 7, pp. 1195–1203, 1994.

[17] M. El Komy, Y. Abdelrahman, M. Funk, T. Dingler, A. Schmidt, and S. Abdennadher, "ABBAS: An adaptive bio-sensors based assistive system," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, May 2017, pp. 2543–2550.

[18] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (GSR) as an index of cognitive load," in *Proc. CHI Extended Abstracts Hum. Factors Comput. Syst.*, 2007, pp. 2651–2656.

[19] M. Bilalpur, M. Kankanhalli, S. Winkler, and R. Subramanian, "EEG-based evaluation of cognitive workload induced by acoustic parameters for data sonification," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 315–323.

[20] D. Dearing, A. Novstrup, and T. Goan, "Assessing workload in human-machine teams from psychophysiological data with sparse ground truth," in *Communications in Computer and Information Science*. Springer, 2019, pp. 13–22, doi: 10.1007/978-3-030-14273-5_2.

[21] S. Chen and J. Epps, "Atomic head movement analysis for wearable four-dimensional task load recognition," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2464–2474, Nov. 2019.

[22] K. Ross, P. Sarkar, D. Rodenburg, A. Ruberto, P. Hungler, A. Szulewski, D. Howes, and A. Etemad, "Toward dynamically adaptive simulation: Multimodal classification of user expertise using wearable devices," *Sensors*, vol. 19, no. 19, p. 4270, Oct. 2019.

[23] A. Jimenez-Molina, C. Retamal, and H. Lira, "Using psychophysiological sensors to assess mental workload during Web browsing," *Sensors*, vol. 18, no. 2, p. 458, Feb. 2018.

[24] L. Kohout, M. Butz, and W. Stork, "Using acceleration data for detecting temporary cognitive overload in health care exemplified shown in a pill sorting task," in *Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2019, pp. 20–25.

[25] M. Mazher, A. A. Aziz, A. S. Malik, and H. U. Amin, "An EEG-based cognitive load assessment in multimedia learning using feature extraction and partial directed coherence," *IEEE Access*, vol. 5, pp. 14819–14829, 2017.

[26] T. Matković and V. Pejović, "Wi-mind: Wireless mental effort inference," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Singapore, 2018, pp. 1241–1249.

[27] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Troster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, Mar. 2010.

[28] G. J. K. Novak, K. Stojmenova, and J. Sodnik, "Assessment of cognitive load through biometric monitoring," in *Proc. 7th Int. Conf. Inf. Soc. Technol., Soc. Inf. Syst. Comput. Netw.*, 2017, pp. 303–306.

[29] F. Schaule, J. O. Johanssen, B. Bruegge, and V. Loftness, "Employing consumer wearables to detect office workers' cognitive load for interruption management," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Singapore, vol. 2, 2018, pp. 32:1–32:20.

[30] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J. Biomed. Informat.*, vol. 73, pp. 159–170, Sep. 2017.

[31] M. Gjoreski, M. Z. Gams, M. Lustrek, P. Genc, J.-U. Garbas, and T. Hassan, "Machine learning and end-to-end deep learning for monitoring driver distractions from physiological and visual signals," *IEEE Access*, vol. 8, pp. 70590–70603, 2020.

[32] M. Gjoreski, M. Luštrek, and V. Pejović, "My watch says i'm busy: Inferring cognitive load with low-cost wearables," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 1234–1240.

[33] J. Tervonen, K. Pettersson, and J. Mäntyjärvi, "Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors," *Electronics*, vol. 10, no. 5, p. 613, Mar. 2021.

[34] S. Banerjee, T. Chattopadhyay, A. Pal, and U. Garain, "Automation of feature engineering for IoT analytics," *ACM SIGBED Rev.*, vol. 15, no. 2, pp. 24–30, 2018.

[35] M. Dietz, I. Aslan, D. Schiller, S. Flutura, A. Steinert, R. Klebbe, and E. André, "Stress annotations from older adults–exploring the foundations for mobile ML-based health assistance," in *Proc. 13th EAI Int. Conf. Pervas. Comput. Technol. Healthcare*, May 2019, pp. 149–158.

[36] A. Salfinger, "Deep learning for cognitive load monitoring: A comparative evaluation," in *Proc. Adjunct ACM Int. Joint Conf. Pervasive Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, Sep. 2020, pp. 462–467.

[37] X. Li and M. De Cock, "Cognitive load detection from wrist-band sensors," in *Proc. Adjunct ACM Int. Joint Conf. Pervas. Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, Sep. 2020, pp. 456–461.

[38] S. Barua, M. U. Ahmed, and S. Begum, "Towards intelligent data analytics: A case study in driver cognitive load classification," *Brain Sci.*, vol. 10, no. 8, p. 526, Aug. 2020.

[39] Y. Yoshida, H. Ohwada, F. Mizoguchi, and H. Iwasaki, "Classifying cognitive load and driving situation with machine learning," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 3, pp. 210–215, Jun. 2014.

[40] A. Yomna, E. Velloso, T. Dingler, A. Schmidt, and F. Vetere, "Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–20, 2017.

[41] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive load estimation in the wild," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–9.

[42] T. Appel, N. Sevcenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, E. Kasneci, and P. Gerjets, "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 154–163.

[43] S. Chen, J. Epps, and F. Chen, "A comparison of four methods for cognitive load measurement," in *Proc. 23rd Austral. Comput.-Hum. Interact. Conf.*, 2011, pp. 76–79.

[44] N. Nourbakhsh, Y, Wang, and F. Chen, "GSR and blink features for cognitive load classification," in *Proc. IFIP Conf. Hum.-Comput. Interact.* Berlin, Germany: Springer, 2013, pp. 159–166.

[45] B. Yin, N. Ruiz, F. Chen, and M. A. Khawaja, "Automatic cognitive load detection from speech features," in *Proc. 19th Australas. Conf. Comput.-Hum. Interact., Entertaining User Interfaces*, 2007, pp. 249–255.

[46] M. V. Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–5.

[47] D. J. McDuff, J. Hernandez, S. Gontarek, and R. W. Picard, "COGCAM: Contact-free measurement of cognitive stress during computer tasks with a digital camera," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 4000–4004.

[48] M. Weiser, "The computer for the 21st century," *Sci. Amer.*, vol. 265, no. 3, pp. 94–105, 1991.

[49] J. B. Carroll, *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[50] J. W. French, R. B. Ekstrom, and L. A. Price, *Manual for kit of Reference Tests for Cognitive Factors*, J. W. French, R. B. Ekstrom and L. A. Price, Eds. Princeton, NJ, USA: Educational Testing Service, 1969.

[51] M. Amboni, L. Iuppariello, A. Iavarone, A. Fasano, R. Palladino, R. Rucco, M. Picillo, I. Lista, P. Varriale, C. Vitale, M. Cesarelli, G. Sorrentino, and P. Barone, "Step length predicts executive dysfunction in Parkinson's disease: A 3-year prospective study," *J. Neurol.*, vol. 265, no. 10, pp. 2211–2220, Oct. 2018.

[52] M. Liparoti, M. D. Corte, R. Rucco, P. Sorrentino, M. Sparaco, R. Capuano, R. Minino, L. Lavorgna, V. Agosti, G. Sorrentino, and S. Bonavita, "Gait abnormalities in minimally disabled people with multiple sclerosis: A 3D-motion analysis study," *Multiple Sclerosis Rel. Disorders*, vol. 29, pp. 100–107, Apr. 2019.

[53] L. L. Thurstone and T. G. Thurstone, *Factorial Studies of Intelligence* (Psychometric Monograph), vol. 2. 1941, p. 94.

[54] O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *Int. J. Neural Syst.*, vol. 27, no. 2, Dec. 2016, Art. no. 1650041.

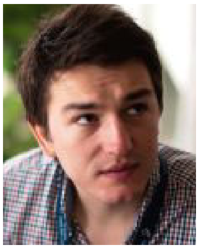
MARTIN GJORESKI received the Ph.D. degree in computer science from the Jožef Stefan Postgraduate School, Ljubljana, Slovenia, in 2020. From 2014 to 2020, he was a Research Assistant with the Department of Intelligent Systems, Jožef Stefan Institute. Since 2021, he has been a Postdoctoral Researcher with the Faculty of Informatics, Università della Svizzera italiana, Switzerland. His research interest includes the development of machine learning methods for monitoring human physical and psychological behavior. Jointly with several research teams, he has received five machine-learning competitions, from 2018 to 2020, including the Sussex-Huawei Locomotion Challenge, in 2018 and 2019.

BHARGAVI MAHESH received the M.Sc. degree in autonomous systems from the Bonn-Rhein-Sieg University of Applied Sciences, in 2019. Since then, she has been a part of the Fraunhofer Institute of Intelligent Circuits, Erlangen, as a Research Assistant. She is currently a Research Associate under Smart Sensing Electronics Division. Her research interests include applications of machine learning algorithms in areas, including multi-modal human stress detection and digital sensory perception.

TINE KOLENIK is a Junior Researcher at the Department of Intelligent Systems at the Jožef Stefan Institute and an Assistant Lecturer at the Jožef Stefan International Postgraduate School. His research includes cognitive science, artificial intelligence, persuasive technology, digital mental health, behavioral public policy, behavioral big data, natural and artificial cognitive architectures, philosophy of science, and 5E cognition. His Ph.D. is on the intelligent cognitive assistant technology for behavior change in mental health. Kolenik is also an assistant editor for the international journal of computing and informatics Informatica and Center for Cognitive Science of University of Ljubljana collaborator.

JENS UWE-GARBAS received the Diploma and Ph.D. degrees in electrical engineering from Friedrich-Alexander-University Erlangen-Nuremberg, in 2004 and 2010, respectively. In 2010 he joined the Fraunhofer-Institute for Integrated Circuits IIS. He was appointed as the Head of the Intelligent Systems Group, in 2011, and the Deputy Head of the Electronic Imaging Department, in 2012. Since 2021, he has been the Co-Director of the Smart Sensing and Electronics Division. His research interests include video coding, affective computing, and automated facial analysis.

DOMINIK SEUSS received the M.Sc. degree in applied computer science from the University of Bamberg, Germany, where he is currently pursuing the Ph.D. degree. After graduation, he joined Fraunhofer IIS, Erlangen, Germany, where he continued his research on automatic facial expression analysis. Since 2019, he has been the Head of the Intelligent Systems Group and responsible for industrial and public research projects as well as software licensing in the areas of real-time computer vision, affective computing, and facial analysis. His research interests include machine learning topics in the field of deep learning in computer vision and generative networks.



HRISTIJAN GJORESKI received the M.Sc. and Ph.D. degrees in information and communication technologies from the Jozef Stefan Postgraduate School, Slovenia, in 2011 and 2015, respectively. From 2010 to 2016, he was a Researcher with the Department of Intelligent Systems, Jozef Stefan Institute, Slovenia. In 2017, he was a Postdoctoral Research Fellow with the University of Sussex, U.K. He is currently an Assistant Professor with the Ss. Cyril and Methodius University, Skopje, North Macedonia. His research interests include artificial intelligence, machine learning, and wearable computing. He was highly successful at several machine learning competitions. He received the First Place Award at the EvAAL Activity Recognition Challenge, in 2013, the ChallengeUP Fall Detection Competition, in 2019, and the Emteq Activity Recognition Challenge at Ubicomp, London, U.K., in 2019.

MITJA LUŠTREK received the Ph.D. degree from the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, in 2007. He held a postdoctoral position with the Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock, Germany, in 2010. He has worked with the Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia, ever since. He was a principal investigator with a number of international research projects on this topic. He is currently the Head of the Ambient Intelligence Group. His research interests include analysis of sensor and other data related to human health and behavior using machine learning. He was highly successful at several computer science competitions, such as the XPrize Tricorder competition, EvAAL competition and Sussex-Huawei Locomotion Challenge 2018–2020. He also served as the Chair for the Slovenian Artificial Intelligence Society for two terms.

MATJAŽ GAMS (Member, IEEE) received the Ph.D. degree. He is currently the Head of the Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, and a Professor of computer science with the University of Ljubljana and the Jozef Stefan Postgraduate School. His professional interests include intelligent systems, artificial intelligence, cognitive science, intelligent agents, electronic and mobile health, business intelligence, and information society. He is a member of several international program committees of scientific meetings, national and European strategic boards and institutions, editorial boards of 11 journals, and the Managing Director of the journal *Informatica*. His team won two activity recognition competitions and placed in the finals of the XPrize Tricorder Competition. He is also a member of the National Council of Slovenia, representing the field of science for the term, from 2017 to 2022.

VELJKO PEJOVIĆ received the Ph.D. degree in computer science from the University of California at Santa Barbara, USA. Since 2015, he has been an Assistant Professor with the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. Prior to this, he was a Research Fellow with the Department of Computer Science, University of Birmingham, U.K. His research interests include mobile computing, HCI, and resource-efficient computing. His work on mobile interruptiblity won the Best Paper Nomination at ACM UbiCompc 2014, while his work on epidemics modeling won the Orange D4D challenge, in 2013. More about his research can be found at http://lrss.fri.uni-lj.si/Veljko/