Journal of Biomedical Informatics 139 (2023) 104299

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

An interpretable machine learning approach to multimodal stress detection in a simulated office environment

Mara Naegelin [a,b,*,1], Raphael P. Weibel [a,b,1], Jasmine I. Kerr [a,b,1], Victor R. Schinazi [c,d], Roberto La Marca [e,f], Florian von Wangenheim [b,d], Christoph Hoelscher [d,g], Andrea Ferrario [a,b]

[a] Mobiliar Lab for Analytics at ETH Zurich, Department of Management, Economics, and Technology, ETH Zurich, Weinbergstrasse 56/58, Zurich, 8092, Switzerland
[b] Chair of Technology Marketing, Department of Management, Economics, and Technology, ETH Zurich, Weinbergstrasse 56/58, Zurich, 8092, Switzerland
[c] Department of Psychology, Bond University, 14 University Drive, Robina, 4226, Australia
[d] Future Health Technologies, Singapore-ETH Centre, 1 Create Way, Singapore, 138602, Singapore
[e] Centre for Stress-Related Disorders, Clinica Holistica Engiadina, Plaz 40, Susch, 7542, Switzerland
[f] Chair of Clinical Psychology and Psychotherapy, Department of Psychology, University of Zurich, Binzmuehlestrasse 14, Zurich, 8050, Switzerland
[g] Chair of Cognitive Science, Department of Humanities, Social and Political Sciences, ETH Zurich, Clausiusstrasse 59, Zurich, 8092, Switzerland

ARTICLE INFO

Keywords:
Stress detection
Machine learning
Mouse movements
Keystroke dynamics
Heart rate variability
Interpretability

ABSTRACT

Background and objective: Work-related stress affects a large part of today's workforce and is known to have detrimental effects on physical and mental health. Continuous and unobtrusive stress detection may help prevent and reduce stress by providing personalised feedback and allowing for the development of just-in-time adaptive health interventions for stress management. Previous studies on stress detection in work environments have often struggled to adequately reflect real-world conditions in controlled laboratory experiments. To close this gap, in this paper, we present a machine learning methodology for stress detection based on multimodal data collected from unobtrusive sources in an experiment simulating a realistic group office environment (N=90).

Methods: We derive mouse, keyboard and heart rate variability features to detect three levels of perceived stress, valence and arousal with support vector machines, random forests and gradient boosting models using 10-fold cross-validation. We interpret the contributions of features to the model predictions with SHapley Additive exPlanations (SHAP) value plots.

Results: The gradient boosting models based on mouse and keyboard features obtained the highest average F1 scores of 0.625, 0.631 and 0.775 for the multiclass prediction of perceived stress, arousal and valence, respectively. Our results indicate that the combination of mouse and keyboard features may be better suited to detect stress in office environments than heart rate variability, despite physiological signal-based stress detection being more established in theory and research. The analysis of SHAP value plots shows that specific mouse movement and typing behaviours may characterise different levels of stress.

Conclusions: Our study fills different methodological gaps in the research on the automated detection of stress in office environments, such as approximating real-life conditions in a laboratory and combining physiological and behavioural data sources. Implications for field studies on personalised, interpretable ML-based systems for the real-time detection of stress in real office environments are also discussed.

1. Introduction

Since the mid-twentieth century, the nature of the workplace has changed dramatically [1,2]. These changes include the transition to knowledge work, the pervasive use of information and communication technology, and the need for a more flexible workforce [3]. At the same time, the prevalence of work-related stress has been increasing [4,5] and is reported to be the most frequent mental health issue at the workplace [6]. More recently, the global COVID-19 pandemic with its dramatic impact on work and social environments has led to an alarming rise of mental stress and related depressive symptoms [7,8]. These

* Corresponding author. Mobiliar Lab for Analytics at ETH Zurich, Department of Management, Economics, and Technology, ETH Zurich, Weinbergstrasse 56/58, Zurich, 8092, Switzerland.
E-mail address: mnaegelin@ethz.ch (M. Naegelin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jbi.2023.104299
Received 29 July 2022; Received in revised form 16 December 2022; Accepted 26 January 2023
Available online 30 January 2023

developments call for the effective detection of acute work-related stress to avert the onset of stress-related symptoms, physical diseases and mental disorders resulting from the chronification of stress [9–11]. Indeed, the continuous detection of stress may support self-monitoring, self-awareness and self-management of one's stress levels [12,13]. Advances in sensor technologies and the progress in computational power and concurrent rise of machine learning (ML) methods in the last decades have made the automatic and data-driven detection of stress in real life attainable. Moreover, continuous stress detection allows for the design of a specific kind of tailored digital health intervention called just-in-time adaptive intervention (JITAI) [14,15]. In a JITAI, selected components, such as intervention prompts, may be adapted and personalised to an individual's context and changing stress levels in real time [16,17]. In addition to increasing adherence [18], adaptive, personalised intervention prompts help reduce symptoms and change or increase desired health-related behaviour [19–21].

To achieve continuous, personalised, and real-time stress detection, it is critical that researchers conduct studies which enable the development of prediction models for stress based on unobtrusively collectable data. A wide range of measurable physiological or behavioural indicators of the stress response have been investigated for this purpose [6,22,23]. Some physiological signals (e.g., heart rate variability, electrodermal activity, skin temperature) may be collected through wearable biosensors, which are becoming increasingly comfortable, affordable and accurate [24]. In addition, environmental sensors and integrated software solutions have been used to collect contextual (e.g., ambient sound, location, calendar entries) and behavioural data (e.g., mobile phone usage, facial expressions, keystrokes, mouse movements) related to stress [15,22,25]. To process the high volumes of data collected in stress detection studies from multimodal sources, researchers have come to rely on data-driven ML algorithms [6,26].

Although a considerable number of stress detection studies have been conducted in the last two decades [for reviews, see, e.g., 15,22,23,27], and despite the progress in wearable sensor technology and the rise of ML methods, there remain a number of challenges regarding the detection of stress in real-world contexts. Consequently, researchers struggle to adequately translate—perhaps overly optimistic—results from controlled laboratory experiments into unconstrained environments, such as offices [28]. Specifically, we argue that the following six challenges regarding the automated detection of stress levels in a laboratory environment need to be overcome.

First, previous studies have usually relied on artificial stress-elicitation methods (e.g., mental arithmetic tasks, the Stroop test, or the cold pressor test), tasks and experiment protocols that do not properly reflect real-life contexts, such as an office environment [23]. Second, despite strong evidence that relying on data from multiple sources to predict stress increases the predictive power of stress detection models [22], many combinations of different data sources (e.g., physiological cardiac activity data together with behavioural computer interaction data) have yet to be studied in detail. Third, the derivation of adequate ground truth labels is necessary for the detection of stress. However, in stress detection experiments, labels have often been derived from the assigned experimental conditions [23], while in the case of office environments and other real-world contexts, this kind of information is not available. Fewer studies have used psychological measures, usually based on self-reports [29,30]. In addition, multiclass or regression models for stress detection are still uncommon compared to binary classification (i.e., detecting "stress" versus "no stress"), even though an increased granularity might more adequately reflect the reality of the stress response [26]. Fourth, the majority of studies have featured single-participant settings with relatively small sample sizes [31]. As a result, there is a scarcity of more realistic group scenarios, while the number of data points available for training of the ML algorithms is also limited. Fifth, on the methodological side, the relatively small size of generated data sets limits the choice of ML algorithms to be used to detect stress [32,33]. It follows that research has yet to leverage the full potential of more complex (and data-intensive) ML algorithms, such as boosting and deep neural nets, especially in combination with extensive hyperparameter tuning [26,34]. Finally, beyond the difficulties of moving from laboratory to field environments, additional challenges will have to be solved in order to integrate a developed stress detection model into a JITAI. For one, based on granular information of the stress level over time, an intervention prompt logic has to be defined in such a way that the intended positive effect of the health intervention is maximised [35,36]. For another, stress level-feedback based on the predictions of a developed ML model, which may be provided to users to improve their stress self-management, should ideally allow them to make autonomous, informed decisions regarding their health. To that end, users should be able to understand this feedback, fostering their autonomy and trust in the JITAI [37,38], and potentially increasing the adherence to the stress management programme. Despite the growing body of literature on ML interpretability methods, also known as "explanations" [39,40], these methods have rarely been applied to stress detection models [14,41].

With this work we thus pursue the following research goal: to design a methodology to address the aforementioned six challenges in detecting stress levels in a laboratory environment and provide practical implications for automated stress detection in real offices. To do so, we have developed a methodology for the detection of self-assessed perceived stress, valence and arousal levels with ML models based on multimodal data from a laboratory experiment where we simulated a realistic group office setting in a controlled environment and exposed participants to different work-related stressors.

Specifically, a total of 90 participants in three experimental conditions were tasked with basic workload throughout the experiment, while intermittently being subjected to work interruptions and social pressure. Social pressure was induced with an adapted version of the Trier Social Stress Test for Groups (TSST-G), a standardised and well-validated psychosocial stress test [42]. We have previously described this experimental protocol in our manuscript [43], in which we examined group differences in the psychobiological stress response (e.g., salivary cortisol) over time and explored the role of cognitive stress appraisal as a mediator of the response, in order to contribute to a better understanding of the psychophysiological processes underlying work-related stress. Importantly, the results in [43] provide proof of a successful stress elicitation through our experimental protocol and thus justify the use of the data set for the purpose of the current manuscript.

Hence, the methodology introduced in this work is developed on the collected *behavioural* mouse and keyboard data, *physiological* heart rate variability data, and *psychological* self-reported multidimensional stress level data consisting of perceived stress, valence and arousal. Its contributions in relation to the state-of-the-art include the combination of different behavioural and physiological data modalities, multiclass ML modelling using support vector machines, random forests and gradient boosting algorithms, implementing extensive hyperparameter tuning, and using Synthetic Minority Oversampling Technique (SMOTE) [44] to cope with multiclass imbalance.

Finally, to investigate and promote the interpretability of our ML model results, our methodology comprises the use of different SHapley Additive exPlanations (SHAP) value plots [45] to derive explanations for features and feature interactions characterising different levels of perceived stress, valence and arousal.

The remainder of this paper is structured as follows. Section 2 provides an overview of the psychophysiology of stress and the measurement of the different stress responses. Section 3 summarises the related works from automated stress detection research, while discussing the limitations of existing laboratory studies and the contributions of our work in more detail. Section 4 describes the data collection, feature generation and modelling steps of our methodology. Results are provided in Section 5 and their implications discussed in Section 6. Section 7 contains the conclusion and directions for future research.

2. Background

The transactional model of stress by Lazarus and Folkman [46] proposes that stress is a consequence of interactions between an organism and its environment. Appraising a stressor as personally relevant and the disposable resources as insufficient triggers a stress response. These psychological and multi-system physiological responses elicited by a stressor [47] can also result in subconscious and intentional behavioural changes [23]. Prolonged or repeated exposure to acute stressors leads to chronic stress, which in turn has been linked to a wide range of negative health outcomes including burnout and depression [48], and cardiovascular diseases [11]. For continuous stress detection, researchers try to capture acute stress events using quantifiable markers of the different stress responses. The monitoring of daily acute stress can then enable stress management JITAIs to provide targeted, personalised support, with the aim to prevent the development of chronic stress and the related negative health consequences [49].

2.1. Psychological stress response

Psychological processes are an integral part of the stress response. They have in some cases been used as a basis to establish ground truth, that is to provide labels on which to train ML algorithms for stress detection [23]. Stressors—even standardised laboratory stress tests—may elicit a wide range of emotions and thoughts [47]. Therefore, acute stress may be captured through several components of the psychological response. For example, questionnaires that assess degrees of perceived stress, valence and arousal are reflective of acute stress [50]. Technology has facilitated the collection of momentary stress experiences and events throughout the day (i.e., Ecological Momentary Assessment, EMA [16]) in real-world environments, such as the workplace. Nevertheless, self-reporting requires the active and repeated contribution of a participant and can only yield discrete measurements.

2.2. Physiological stress response

In addition to the psychological response, experiencing stress also results in an activation of the two major physiological stress systems: the hypothalamic-pituitary-adrenal axis and the autonomic nervous system (ANS). A wide range of biomarkers of these systems have been used to measure stress [23]. However, some of these markers can be cumbersome to collect (e.g., cortisol levels from saliva or blood samples, electroencephalogram [EEG] via electrode cap) and are therefore unsuitable for unobtrusive, continuous stress detection [15]. Other physiological signals, such as cardiac or electrodermal activity, can be collected non-invasively via external sensors and are therefore studied more frequently [26]. Particularly popular are measures of heart rate variability (HRV), which quantifies the variability in inter-beat-interval durations and is sensitive to ANS activity alterations associated with stress [51]. High HRV is generally linked to improved health and social functioning, while low HRV is associated with a wide range of disorders and diseases [52].

HRV may be quantified by a variety of time domain, frequency domain, and non-linear measures, many of which have been consistently and successfully used to distinguish between stressful and relaxed states [53]. With recent developments in wearable technology, HRV data collection has become less obtrusive and cardiac data can now be acquired via wearable electrode holsters, chest straps, or—though still at somewhat lower quality—with photoplethysmogram (PPG) sensors in arm- and wrist-worn devices [24]. Nevertheless, the collection of cardiac data does require the use of a sensor device, which may be disagreeable to some potential users. The signal is also prone to distortion by motion artefacts and affected by confounds such as physical activity [54].

2.3. Behavioural stress response

Behavioural changes induced by the psychophysiological stress response may be reflected in changes in body posture, facial expression, or one's interaction with the environment [22]. While behavioural measurements for stress detection have not been studied as extensively as physiological ones [26], initial studies provide evidence for their predictive power for stress detection, and additional research may further strengthen these results [22]. In addition, many behavioural measures can be collected non-invasively and without the need for extra equipment (e.g., through video or audio data collected from sensors integrated in smartphones or laptops), making them both more affordable and less obtrusive than physiological signals. However, data sources such as cameras and microphones might be too privacy-invading for potential users, especially in a work context [15,55].

In this regard, mouse and keyboard data have been suggested as some of the best-suited modalities for stress detection in a work environment [15,32]. While an inherent drawback of mouse and keyboard data is that they are only generated when one interacts with a computer, most office employees spend a large part of their day in front of their computer [3]. The association of mouse and keyboard-related measures with stress can be theoretically grounded in the neuromotor noise theory by van Gemmert and van Galen [56]. The theory suggests that the imbalance between resources and demands leads to an increased signal-to-noise ratio in the brain, which is reflected in increased variations in human movements. Indeed, mouse and keyboard data have successfully been used in emotion recognition studies [57–60]. However, some researchers have raised doubts as to whether these data sources have enough correlation with stress to be used for its detection [61,62].

The achievement of a consensus on the predictive potential of keyboard and mouse features for stress detection is impeded by the lack of consistent, theoretically grounded frameworks for feature derivation for either data modality [63]. Regarding keyboard data, many studies draw from the early works of Vizer et al. [59] and Epp et al. [60] to derive a range of features related to typing behaviour, for example regarding key counts and keystroke dynamics. For mouse data, there are no clear reference works, thus researchers often develop their own features based on different characterisations of mouse usage and mouse movements in particular [57,62].

3. Related work

In this section we discuss recent works on automated stress detection with a focus on studies intending to simulate work-related stress in office environments. For the convenience of the reader, the most relevant related works have been summarised in Table 1 in terms of collected data modalities, characteristics of the experimental protocol, ML modelling and use of interpretability methods.

3.1. Multimodal stress detection

Although the different psychobiological stress responses are often closely associated with one another, they do not always align [50]. Therefore, researchers have frequently collected data from multiple modalities and found that the combination of different modalities generally improves the performance of stress detection models [29,67,70]. However, instead of combining as many data sources as possible, the optimal choice of modalities to be used in a stress detection model should find a balance between prediction performance and other important evaluation criteria. Indeed, not all potential data sources are appropriate for stress detection in a work context such as an office environment. Besides enabling reliable and continuous stress detection, data collection should not obstruct employees' work, interfere with their daily routines or pose a potential threat to their privacy [15]. Here, we argue that mouse and keyboard data are among the least

Table 1
Overview of related works on automated stress detection in laboratory environments (ordered alphabetically by first author).

	Data modalities		Experiment protocol				ML modelling					iML[g]
	Physiological	Behavioural	Work[a]	Stressors[b]	Baseline task	Label derivation	Cl.[c]	N[d]	n[e]	Algorithms	Tune[f]	
[64]	EDA, EMG, PPG, RSP, ST	posture, FE	–	MAT (6 intensities)	resting	(1) self-report, (2) observer assessment, (3) physiology assessment	3	25	126	SVM	✓	–
[65]	ECG, EDA, ST	–	✓	adapted TSST, adapted SCWT, MAT	resting	baseline and stressor phases	2	34	n.a.	LDA, QDA, SVM, KNN	–	–
[66]	ECG, EDA, EEG	–	✓	MAST (CPT, MAT)	resting	baseline and stressor phases	2	15	24	SVM	–	–
[32]	–	M, K	✓	M: SCWT (2 intensities), K: MAT	M task: questionnaire, K task: report writing	M: intensity of SCWT before M task, K: presence of MAT during K task	2	25	M: 174, K: 188	KNN, SVM, NB	–	–
[30][h]	PPG, RSP	–	–	public speaking, MAT, CPT	resting	(1) self-report, (2) baseline and stressor phases	2	32	1700	SVM, RF, Adaboost, GB, LR	✓	–
[62]	–	M	–	MAT	game-like M tasks	intensity of MAT before M task	2	53	86–106	LR, KNN, SVM, RF	✓	–
[67][h]	PPG, EDA, ST	–	–	MAT (3 intensities)	resting	baseline phase, low & high stress from self-report	3	21	n.a.	DT, NB, KNN, SVM, bagging, GB, RF	–	–
[33]	ECG, RSP	–	✓	MIST (MAT in 2 intensities)	resting	baseline phase & MAT intensity	3	39	702	SVM, LDA, KNN, Adaboost	✓	–
[41][i]	ECG, EDA, EMG, RSP, ST	–	–	TSST	resting	baseline and stressor phases	2	15	1640	RF	✓	✓
[29]	ECG, EDA	posture, FE, M, K, app log	✓	time pressure, interruptions	report writing	(1) baseline and stressor phases, (2) self-report	2, regr.	25	2688	NB, KNN, BN, DT, MLP, SVM, LinR	–	–
[68][i]	ECG, EDA	–	–	TSST	resting	self-report	regr.	15	1353	RF, GB, RNN	✓	–
[61]	–	M, K	✓	2 intensities	M: game-like tasks, K: typing task	self-report	3	62	M: 411, K: 429	RF-based MIL	–	–
[63]	–	M, K	–	3 intensities	essay writing	self-report	3	41	123	SVM, RF, NB, DT	–	–
[34]	EDA, PPG, ST	FE, speech, K	–	visual stimuli	–	self-report	2	40	2400	RNN	✓	–

Abbreviations: ECG = electrocardiogram; EDA = electrodermal activity; EEG = electroencephalogram; EMG = electromyogram; PPG = photoplethysmogram; RSP = respiration; ST = skin temperature; FE = facial expressions; K = keyboard; M = mouse; CPT = cold pressor test; MAST = Maastricht Acute Stress Test; MAT = mental arithmetic task; MIST = Montreal Imaging Stress Task; SCWT = Stroop Color Word Test; TSST = Trier Social Stress Test; BN = Bayes net; GB = gradient boosting; KNN = K-nearest-neighbours; DT = decision tree; LDA = linear discriminant analysis; LinR = linear regression; LR = logistic regression; MIL = multiple instance learning; MLP = multilayer perceptron; NB = naïve Bayes; QDA = quadratic discriminant analysis; RF = random forest; RNN = recurrent neural net; SVM = support vector machine.

[a]Intending to simulate work-related stress.

[b]Intensity of tasks is usually varied via cognitive load, time pressure, and/or socio-evaluative threat.

[c]Number of classes.

[d]Number of participants.

[e]Total number of observations.

[f]Mention a hyperparameter tuning strategy.

[g]Apply interpretable machine learning methods.

[h]Lab study only.

[i]Use the WESAD data set [69].

obtrusive modalities and have enormous potential for scalability in office environments, while current wearables can enable the continuous and flexible collection of cardiac data.

In previous works, HRV features have frequently been used in combination with other physiological signals such as electrodermal activity [cf., e.g., 66,71]. Despite similar collection requirements, very few studies have combined both mouse and keyboard data [63], although some authors have developed separate models on separate data sets for each of these modalities [32,61]. The combination of data across both physiological and behavioural modalities is less common. For example, in [34], the authors used a wrist-worn device to capture cardiac, electrodermal and skin temperature data, and a smartphone to collect facial expression, speech and touch features. To the best of our knowledge, the laboratory experiment described in [29] is the only one where mouse, keyboard and HRV data have been collected conjointly, among additional data sources such as facial expressions and body posture collected using video cameras.

3.2. Experimental protocols of stress detection studies

The majority of existing studies on stress detection have relied on laboratory experiments where stress is induced using artificial stressors [30,33,64,66]. To this end, the studies usually followed a similar protocol in which a non-stressed baseline phase was followed by periods of stress elicitation, interspersed and/or succeeded by a resting phase. In some studies, the validity of the protocol to elicit stress-related changes in the respective phases was assessed with statistical comparisons of intermittently sampled questionnaire [30,32,62] or—more rarely—biochemical data [65,66].

The experimental protocol is commonly used to label the collected data into periods of "stress" and "no stress" [29,32,62], but this kind of ground truth does not exist in field studies. Assigning all participants the same level of stress based on the presence or absence of a stressor might also not adequately address inter-personal differences in stress reactivity [47]. Thus, some studies have also relied on questionnaire data to derive labels instead, either discretising scales into classes or in regression models [29,61,64].

The chosen stress elicitation task differs widely from study to study. Although some studies have developed their own stress elicitation protocols [29,61,64], the majority of studies relied on adaptations of well-validated stress tests, such as the Stroop test, mental arithmetic tasks, the cold pressor test, and the TSST [for a review, see, e.g., 23]. While such validated stress elicitation procedures are known to evoke a strong stress response for most participants, they are very different from stressful situations that one might experience in the real world. Studies that have tried to mimic real-world stress situations more closely by developing their own stress elicitation protocols often rely on time pressure as the stressor [29,61,65]. However, in real work contexts, time pressure might easily be confounded with a period of increased productivity and cannot necessarily be equated with stress [32].

During baseline or resting experiment phases participants often performed a task that is intended to induce a state of relaxation such as reading magazine articles or watching videos [65,69]. However, in a real-life working scenario, people are continuously executing tasks throughout the day without necessarily experiencing stress. Thus, a relaxation-inducing baseline or resting phase would not reflect this baseline workload accurately. In contrast, studies involving mouse and keyboard data, require a computer task to allow for data collection during stressful and resting phases. Here, researchers have typically designed highly specific tasks which are too artificial and simplistic to simulate real working conditions—for example point-and-click exercises, the Towers of Hanoi game, or queue-prompted text writing [61–63]. In such studies, different experimental states were then evoked by varying the demands of the tasks themselves [29,61], or by periods of stress elicitation immediately preceding the data collection task [32,62].

In summary, despite some of the above mentioned studies specifically aiming to reflect office-like conditions [29,32,61,66], there is still room for improvement regarding the realism of the data collection tasks, baseline workload and stress-elicitation methods in stress detection experiments simulating office environments. In addition, none of these laboratory studies accounted for the social component of shared offices in their single-participant setups, further limiting the generalisability of procedures and results to real-life working scenarios.

3.3. Machine learning-based stress detection

To build a stress detection model researchers usually train ML algorithms on a generated data set [29,30,32,64]. To do so, they derive different features from raw data collected from multiple sources, and use these as input in a binary classification problem aiming to discriminate between a stressed and non-stressed class. However, only a few studies have increased the granularity of their labels to a multiclass classification problem [61,63,67], or considered the collected self-reported stress levels on a Likert scale as regression labels [29,68].

ML algorithms commonly used to detect stress are support vector machines, k-nearest neighbours, and random forests [for reviews, see, e.g., 6,26]. More complex methods, such as boosting, feed-forward and recurrent neural networks have received some attention only in recent years [34,68,70], but the application of deep learning methods is limited by the small size of data sets generated in laboratory experiments. Additionally, studies aiming to detect stress with ML algorithms show a high degree of variability in terms of methods, performance assessment and reporting, which affects the reproducibility and comparability of results [30,33,72]. In fact, apart from the use of different experimental protocols to collect data for ML modelling, different data normalisation procedures are performed [64,66]. Moreover, the reporting of ML results does not often mention model hyperparameters (e.g., the depth of the trees in a random forest ensemble) and their tuning [29,63]. Finally, the measures used to select best-performing models vary among studies, as well. Common choices are the accuracy, the Area Under the ROC Curve (AUC), and the F1 score [30,32,61].

3.4. Interpretable stress detection models

ML algorithms are capable of capturing complex high-dimensional relationships in data, but their complexity means that the underlying decision rules are often not easily understood. This low transparency may become disadvantageous for a stress detection system based on nonlinear models for two reasons. First, it impedes the efficient debugging and validation of ML models for stress detection during their design process. Second, it affects the implementation of ML methods to detect stress as humans are less willing to trust, adopt and appropriately use systems that are not understandable and transparent [38,39]. To overcome these issues, researchers have introduced different interpretable Machine Learning (iML) techniques with the aim to increase human understanding of and to foster trust in opaque ML methods [39,40]. One common iML approach is feature attribution, where a weight is estimated for each feature of a model depending on its importance for the model's predictions. For example, tree-based algorithms such as random forests and gradient boosting allow computing the "global" importance of a feature used to train the tree ensembles. These feature importance scores return the numbers of times a feature is used in the ensemble, or a function of the gains of the tree splits using that feature. However, these scores do not provide information on how different features may contribute to the ML model prediction of any given data point, and the different strategies to compute them may lead to inconsistent results [45].

Among all iML methods that can support the understanding of ML model outcomes, SHapley Additive exPlanations (SHAP) [45] are one of the most commonly used in applications. Recent work on affective computing has seen the use of SHAP values to explain dropouts in digital health interventions programmes [73] or to evaluate the feasibility of ML methods to detect emotional states and transitions [74]. In previous research on stress detection and affective computing in general, feature importance methods have in some cases been used as a feature selection technique [33,75] or to identify the data modalities having the most discriminative potential [e.g., 61]. To the best of our knowledge, more advanced methods than off-the-shelf tree-based feature importance techniques, such as SHAP values, have rarely been systematically employed in stress detection research [14,41,76]. Jaber et al. [41] applied SHAP value analysis to the publicly available WESAD data set [69] to design a prototype stress prediction report for clinicians reminiscent of blood test results. Rozet et al. [14] used SHAP-like visualisations of layer-wise relevance propagation scores to interpret the predictions of their stress detection model based on environmental and physical activity field data. Lastly, Ng et al. [76] have recently applied SHAP values to investigate the most predictive features when predicting next-day stress of pregnant women in a proof-of-concept study. However, the full potential of interpretability methods to support engineers in the design of ML models for stress detection and to increase end-user understanding and trust in these models is yet to be harnessed.

4. Methods

In this section, we introduce the key steps of our proposed methodology for the automated detection of stress measures with ML models. In Sections 4.1–4.3 below we briefly describe the laboratory experiment which generated the data set used in this work. Additional details on the experiment procedure can be found in our manuscript [43] in which we first introduced this protocol and confirmed the successful stress elicitation while investigating the psychophysiological processes underlying work-related stress. Then, in Sections 4.4 and 4.5 we detail the data collection and preprocessing steps. Finally, in Sections 4.6 and 4.7 we describe the ML pipeline for the automated detection of stress measures in detail and the method we considered to interpret ML model results. See Fig. 1 for an overview of the proposed methodology.

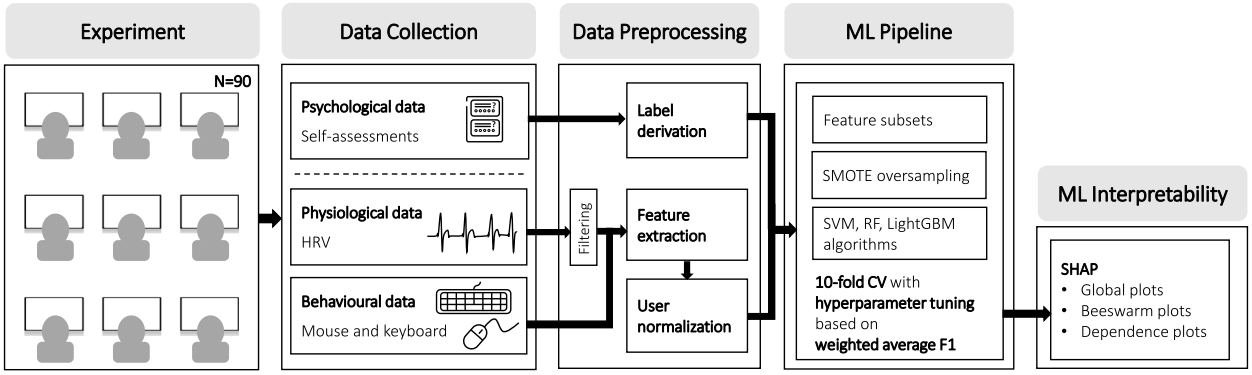


Fig. 1. The different steps of our methodology for the automated detection of self-assessed perceived stress, valence and arousal levels with ML models. We describe the step "Experiment" in Sections 4.1–4.3, while Sections 4.4 and 4.5 are devoted to "Data Collection" and "Data Preprocessing", respectively. Finally, in Sections 4.6 and 4.7 we discuss the "ML Pipeline" and "ML Interpretability" steps. Abbreviations: HRV = heart rate variability; SMOTE = synthetic minority oversampling technique; SVM = support vector machine; RF = random forest; LightGBM = light gradient boosting machine; CV = cross-validation; SHAP = SHapley Additive exPlanations.

4.1. Participants

Ninety participants (44 female) with a mean age of 23.11 (SD=3.80) were recruited via the university's online recruitment website. For a detailed description of exclusion criteria, see [43]. Participants gave written informed consent and were compensated for participation with 75 Swiss francs. The study was approved by the ETH Zurich's ethics commission (EK 2019-N-34).

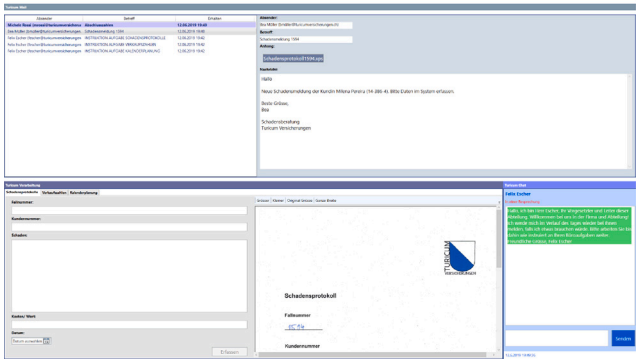
4.2. Environment

ETH Zurich's Decision Science Laboratory was modified to mimic a real-world group office environment. Each participant was assigned a desk equipped with a computer, mouse and keyboard. A custom-built software guided participants through the experiment, displaying instructions screens, questionnaires and the experimental tasks (see Fig. 2). The software was synchronised between all participants and programmed specifically for each experimental condition.

4.3. Procedure

Data were collected during three sessions of 30 participants each. In each session, participants were randomly assigned to one of three experimental conditions (stress condition 1, stress condition 2, and control condition) and transferred to three separate rooms. Before the experiment started, a set of questionnaires was used to assess sociodemographic and psychological baseline characteristics. The experiment itself consisted of six blocks (see Fig. 3), in which participants were asked to act as employees of a fictitious insurance company. After each block, participants filled out a set of psychological state questionnaires. During all blocks except block 4, participants received email messages asking them to perform typical office clerk tasks using the custom-built software. There were three types of tasks: transcribing handwritten claims reports, aggregating sales numbers and scheduling appointments. During block 4, participants from both stress conditions were confronted with a situation designed to elicit psychosocial stress, whereas the control condition experienced a friendly version of the situation. Specifically, we adapted the TSST-G [42] such that in the stress conditions, two line managers (portrayed by actors) interviewed the participants in front of their peers in search for the best candidate for a promotion. In the control condition, the situation was framed as a non-evaluative professional training which involved reading a work-related dialogue aloud in unison. To induce anticipatory stress in the two stress conditions, we announced the upcoming interaction to the participants 20 min in advance. As an additional stressor, participants in stress condition 2 received frequent chat messages from their manager during blocks 2 and 3 (i.e., before and during anticipation), which interrupted them in their workflow with urgent questions related to their tasks and performance.

(a) Experimental setup



(b) Screenshot of experiment software

Fig. 2. (a) The experimental setup simulating an office environment, including desks, computers and saliva sampling equipment. (b) Screenshot of the experiment software displaying the email inbox, an open tab for one of the three experimental tasks (transcribing scanned insurance claim forms) and the chat window. Cf. also Fig. 1 and A.1 in [43].

4.4. Data collection

Participants' mouse and keyboard usage was captured via a custom-built software deployed on their computers. The recorded computer mouse data consists of timestamps (in ms), the x and y coordinates of the cursor on screen (in pixels) and the type of mouse operation performed (i.e., movement, left, right or middle click, left or right double click, scrolling up or down). Mouse activity was logged at a frequency of 8 Hz. The recorded keyboard data consists of timestamps (in ms) and an identifier of the respective key that was pressed. Cardiac activity was collected with the Firstbeat Bodyguard 2 (Firstbeat Technologies

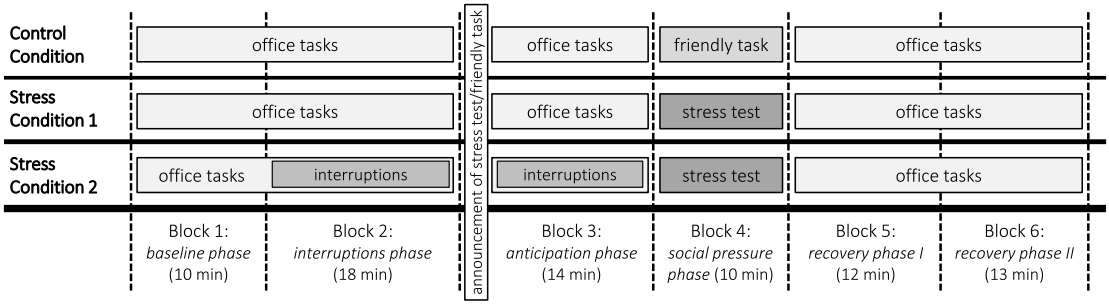


Fig. 3. Study protocol adapted from [43]. Self-assessed psychological measures of stress were collected at the end of each block via questionnaire. Mouse and keyboard activity was logged for all experiment blocks except block 4, where no interaction with the computer took place. Cardiac activity was recorded continuously throughout the experiment.

Oy; Jyväskylä, Finland), an ambulant electrocardiogram device which participants wore throughout the experiment. The device records R-peaks (timestamp and duration; in ms) at a frequency of 1000 Hz. Participants subjective stress load after each block was assessed using the Multidimensional Mood State Questionnaire (MDMQ) [77]. The MDMQ includes the subscales valence ("positive–negative") and arousal ("calm–nervous") with ranges from 4 to 20, where low scores indicate negative valence and nervousness, respectively. A single item on a 5-point Likert scale (1 to 5) was used to assess the perceived level of stress explicitly (i.e., "How stressed are you in this moment?").

In addition to the measures described above, psychological baseline characteristics were measured before the experiment, and additional psychological state questionnaires and biochemical measures (salivary cortisol and salivary alpha-amylase) were collected before the experiment and after each block. For the analysis of these data, see [43].

4.5. Data preprocessing

4.5.1. Measures of stress

From the raw collected mouse, keyboard and cardiac activity data, features were extracted from non-overlapping segments to avoid data leakage. A segment duration of 1 min was chosen, a temporal granularity which has been proposed in previous works [29,54]. In particular, for HRV data, several studies have confirmed that such a segment length is sufficient to distinguish between levels of stress [53,78,79]. However, for mouse and keyboard data, a clear consensus on segment duration has yet to be established, with previous works deriving features on segments ranging from 5 s [61], to one minute [29], to 10 min [32], or even aggregated per given experiment task of varying length [62,63].

The labels assigned to each segment correspond to the self-assessments at the end of the respective experiment block. To reflect the multidimensionality of the psychological stress response, we derived distinct sets of labels from the arousal, valence and perceived stress scales described in Section 4.4. The scales were each divided into three distinct classes to allow for granular predictions, while at the same time removing noise and mitigating class imbalance, similar to [63,80,81]. Specifically, we mapped the scores $\{1,2\}$, $\{3\}$, and $\{4,5\}$ on the perceived stress scale to *low*, *medium* and *high* perceived stress, respectively, the scores $\{4,\dots,9\}$, $\{10,\dots,14\}$, $\{15,\dots,20\}$ on the valence scale to *negative*, *neutral* and *positive* valence, respectively, and the scores $\{15,\dots,20\}$, $\{10,\dots,14\}$, $\{4,\dots,9\}$ on the arousal scale to *low*, *medium* and *high* arousal, respectively.

4.5.2. Mouse features

Mouse features capture either the occurrence count of specific mouse events or the dynamics of mouse movements. Example of occurrence count-related features are the number of clicking or scrolling events within a specific time window, or the number and duration of pauses within the window. We define a mouse movement as a consecutive series of mouse locations on the screen that is delimited by either a click or scroll operation, or a pause of more than 500 ms. The pause threshold was chosen based on previous research [63]. Mouse movements can be characterised by different features. Examples are speed and acceleration, the deviation from an optimal straight line between the start and end point of the movement and the number of direction changes that occurred within a movement. All features extracted in this study were based on features used in previous research [29,57,63,82]. We aggregated all features based on mouse events that could occur more than once within a segment using the mean and standard deviation [57]. If an event did not occur within a segment, the respective features were set to zero. In total, we derived 31 mouse features, see the Appendix for the full list and description of features.

4.5.3. Keyboard features

Keyboard features capture either the occurrence count of specific groups of keys or the dynamics of two or three consecutively pressed keys (so-called di- and trigraphs). A digraph has been defined as two consecutive keys which were recorded at most 500 ms apart. A trigraph is the combination of two consecutive digraphs (sharing the middle key). The 500 ms threshold was chosen based on literature [83] and delimits a pause in the typing behaviour. Examples of occurrence count-related features are the number of character keys pressed during the time window in question, or the number of the delete and backspace key counts. Note that we purposefully did not collect more granular count features which might risk the identification of specific tasks during the experiment. Keystroke dynamics features include the averaged typing speed and pause durations within a segment, the time difference between consecutive keys for digraphs and trigraphs in general, and for the most common specific graphs (i.e., the "ER" digraph and the trigraph with a "Space" key in the middle). All extracted features were based on previous research [29,59,60,63]. Note that our data does not include press and release information on individual keys due to a limitation in the software used to record keyboard activity. As a result, we could not derive some common digraph and trigraph features, such as dwell and flight time [60]. We aggregated all features based on events that could occur more than once within a segment using the mean and standard deviation. If an event did not occur within a segment, the respective features were set to zero, as it has been done in previous research [32]. In total, we derived 19 keyboard features, see the Appendix for the full list and description of features.

4.5.4. HRV features

The Firstbeat Bodyguard 2 device continuously recorded RR interval lengths (in ms) during the experiment. We used the Python package `hrv` (version 0.2.8) [84] to preprocess the RR interval (RRI) data and derive HRV features for each one-minute data segment. RRI time series are subjected to noise caused, for example, by ectopic heart beats or motion artefacts. Filtering and interpolating the RRI tachogram may reduce the impact of such artefacts on the HRV feature estimation [85].

Table 2
Class imbalance of the three stress measures on the full data set and without block 4.

	Level	With block 4	Without block 4
Perceived Stress	*low*	4128 (61%)	3628 (61%)
	medium	1600 (24%)	1400 (24%)
	high	1048 (15%)	868 (15%)
Arousal	*low*	3520 (52%)	3120 (53%)
	medium	2512 (37%)	2162 (37%)
	high	744 (11%)	614 (10%)
Valence	*negative*	142 (2%)	102 (2%)
	neutral	1608 (24%)	1358 (23%)
	positive	5026 (74%)	4436 (75%)

We therefore excluded observations with unrealistic values (under 300 or over 2000 ms) and applied the threshold filter of the `hrv` package to detect outliers. This filter is comparable to the threshold-based algorithm offered by the popular HRV analysis software Kubios [86]. In this approach, each RRI value is compared to the local median and excluded if it differs by more than a certain threshold (chosen as 450 ms or "very low" in our case). All excluded values were replaced by cubic spline interpolations. The filtering results showed that the recordings of two participants were of insufficient data quality (more than 10% non-normal heartbeats), which were therefore excluded from further analysis. From the denoised data, we derived time and frequency domain, and non-linear measures of HRV that are commonly used in literature and that have been shown to reliably discern stressed from non-stressed states on ultra-short recording lengths [53,78]. To calculate the frequency domain features, we estimated the power spectral density of the RRI time series. The estimation was performed via Fast Fourier Transform (using Welch's method with Hann windows) from the detrended and equidistantly re-sampled RRI data (constant trend; cubic spline interpolation at 4 Hz). In total, we derived 9 HRV features, see the Appendix for the full list and description of features.

4.6. Machine learning pipeline

The data preprocessing procedures detailed in Section 4.5 resulted in a data set containing a total of 6776 observations and 59 features derived from three different modalities and 88 participants. In order to adjust for inter-individual differences, we then user-calibrated all non-ordinal features (i.e., all except count-related features) using robust standardisation,

$$x_{\text{standardized}} = \frac{x - \text{Med}}{\text{IQR}},$$

with Med = median and IQR = interquartile range derived from the participant's baseline recording of block 1, similar to [33,63]. Each observation was labelled with the respective level of perceived stress, valence and arousal derived from the subsequent self-report. The measures are affected by different degrees of class imbalance (see Table 2). The preprocessed data set shows that participants mostly perceived *low* levels of stress and arousal, and *positive* levels of valence over the course of the experiment, which is not surprising as stress was only elicited for a subset of participants and experiment blocks. To compare the potential of the behavioural and physiological features for detecting stress, we considered all possible combinations of modalities. Block 4 of our experiment was the only phase where no computer interaction took place (see Fig. 3), hence all mouse and keyboard features for segments of this block have the value zero. As there is no information available for the mouse and keyboard modalities for these observations, we did not use data from block 4 when training classifiers based on only one or both of these two modalities (reducing the number of observations to 5896).

We used support vector machines (SVM), random forests (RF), and light gradient boosting machines (LightGBM) [87] to detect perceived stress, arousal and valence levels separately. LightGBM is a gradient boosting method that employs an ensemble of decision trees where the trees are added in sequence and learned by fitting the negative gradients in each iteration. We used the Python packages `scikit-learn` and `lightgbm` (versions 0.24.1 and 3.2.1, respectively) to implement the ML models. While SVM and RF algorithms have frequently been used in similar studies [e.g., 29,32,62], boosting methods are less common [30,67]. To the best of our knowledge, the novel LightGBM algorithm—which promises higher efficiency and accuracy than other boosting methods [87]—has not yet been applied in stress detection studies. Similar to [29,64], to contextualise our results we compared the best-performing models against a baseline classifier which maximises the class prior (i.e., it always predicts the most frequent class, for each stress measure).

We applied a stratified cross-validation (CV) procedure with 10 folds to tune the hyperparameters of the SVM, RF and LightGBM models, a frequently used approach in affective computing [29,60,63,82,88]. The hyperparameter grids can be found in Appendix. Training data are oversampled using the Synthetic Minority Oversampling TEchnique (SMOTE) [44] to handle the multiclass imbalance, which has yielded promising results in other works [63,65]. The SMOTE algorithm generates synthetic examples of the minority classes along line segments connecting a given data point to its K-nearest neighbours [44]. As a result, for each of the 10 runs of the CV procedure, each ML model was trained on a balanced training data set and evaluated on a test data fold showing multiclass imbalance.

For each feature set, algorithm and hyperparameter configuration, we computed the weighted average precision, recall, F1 score (i.e., the harmonic mean of precision and recall), and AUC across the three classes, for each stress measure, on each of the 10 class-imbalanced test data folds. Precision for each class i is defined as the number of true positives (TP) over the sum of both TP and false positives (FP), recall is defined as the number of TP over the sum of both TP and false negatives (FN), while the F1 score is the harmonic mean of the two, i.e.,

$$\text{precision}_i = \frac{TP_i}{TP_i + FP_i}, \ \text{recall}_i = \frac{TP_i}{TP_i + FN_i},$$
$$\text{F1 score}_i = \frac{2}{\frac{1}{\text{precision}_i} + \frac{1}{\text{recall}_i}}, \ i \in \{0, 1, 2\}.$$

For SVM, we tuned the kernel function, the degree in case of a polynomial kernel, the regularisation parameter, and the kernel coefficient. For RF, we tuned the number of trees in the ensemble, the maximum number of features at each split of a tree in the ensemble, the split criterion, the minimum number of samples at each split, and the maximal depth of each decision tree in the ensemble. In the case of LightGBM, we tuned the boosting type, the number of estimators, the learning rate, the maximal depth of each tree in the ensemble, and the percentage of features at each split of a tree. For each of the three algorithm, we also tuned the number of K-nearest neighbours for the SMOTE oversampling step. We selected the optimal hyperparameter configuration using a grid search optimising the weighted average F1 score over the 10 test data folds as, by definition, the F1 score encodes both the precision and recall of each model [30,33].

4.7. Estimation of SHAP values

We employed SHAP values [45] for our ML model interpretability analyses. SHAP values are scores that are theoretically grounded in the Shapley values used in game theory. The SHAP method decomposes a single prediction of an ML model into the additive contribution of each feature (i.e., the SHAP values). Therefore, the set of SHAP values provides an interpretation of the prediction of any data point (local explanation), unlike other methods such as the off-the-shelf feature importance scores for tree-based ML models (e.g., RF and gradient boosting). In contrast to other feature attribution methods, SHAP values also fulfil a set of desirable properties (i.e., local accuracy, missingness

Table 3
Performance results, averaged over classes and folds, for the best ML models on behavioural features, for all stress measures. Abbreviations: prec = precision; rec = recall; M = mouse features; K = keyboard features.

Algorithm	Features	Perceived stress				Arousal				Valence			
		F1	prec	rec	AUC	F1	prec	rec	AUC	F1	prec	rec	AUC
SVM	M	0.527	0.548	0.515	0.619	0.523	0.534	0.515	0.631	0.705	0.714	0.698	0.670
	K	0.524	0.514	0.545	0.576	0.489	0.504	0.505	0.560	0.663	0.686	0.646	0.617
	MK	0.529	0.557	0.513	0.637	0.531	0.555	0.518	0.645	0.694	0.725	0.676	0.695
RF	M	0.557	0.569	0.550	0.661	0.580	0.594	0.573	0.698	0.749	0.751	0.749	0.737
	K	0.579	0.577	0.582	0.673	0.582	0.583	0.585	0.693	0.751	0.746	0.761	0.741
	MK	0.582	0.585	0.582	0.681	0.580	0.594	0.573	0.698	0.750	0.751	0.750	0.736
LightGBM	M	0.601	0.594	0.612	0.689	0.589	0.586	0.593	0.711	0.752	0.746	0.770	0.743
	K	0.589	0.579	0.615	0.677	0.597	0.594	0.604	0.714	0.756	0.752	0.775	0.755
	MK	**0.625**	**0.618**	**0.643**	**0.729**	**0.631**	**0.630**	**0.635**	**0.757**	**0.775**	**0.772**	**0.782**	**0.796**
Baseline Classifier		0.469	0.379	0.615	0.500	0.366	0.280	0.529	0.500	0.646	0.566	0.752	0.500

Table 4
Performance results, averaged over classes and folds, for the best ML models on HRV features and the combination of HRV with one or both of the behavioural modalities, for all stress measures. Abbreviations: prec = precision; rec = recall; H = HRV features; M = mouse features; K = keyboard features.

Algorithm	Features	Perceived stress				Arousal				Valence			
		F1	prec	rec	AUC	F1	prec	rec	AUC	F1	prec	rec	AUC
SVM	H	0.501	0.484	0.547	0.562	0.503	0.522	0.492	0.626	0.655	0.660	0.651	0.615
	MH	0.531	0.546	0.521	0.629	0.530	0.538	0.524	0.638	0.697	0.710	0.686	0.676
	KH	0.515	0.509	0.523	0.587	0.498	0.511	0.491	0.603	0.658	0.653	0.664	0.587
	MKH	0.534	0.528	0.545	0.575	0.525	0.552	0.511	0.648	0.693	0.693	0.693	0.650
RF	H	0.529	0.537	0.524	0.627	0.531	0.532	0.531	0.640	0.682	0.682	0.684	0.659
	MH	0.580	0.580	0.589	0.686	0.586	0.588	0.586	0.706	0.726	0.726	0.731	0.720
	KH	0.559	0.561	0.557	0.664	0.577	0.578	0.577	0.696	0.733	0.734	0.735	0.740
	MKH	0.589	0.595	0.594	0.702	0.607	0.609	0.606	0.729	0.745	0.745	0.751	0.761
LightGBM	H	0.531	0.536	0.526	0.627	0.528	0.530	0.527	0.644	0.688	0.691	0.685	0.668
	MH	0.599	0.596	0.604	0.708	0.597	0.599	0.597	0.726	0.735	0.734	0.736	0.759
	KH	0.581	0.576	0.588	0.685	0.595	0.594	0.598	0.728	0.745	0.742	0.749	0.771
	MKH	**0.612**	**0.608**	**0.624**	**0.727**	**0.617**	**0.620**	**0.616**	**0.751**	**0.757**	**0.760**	**0.756**	**0.799**
Baseline Classifier		0.461	0.371	0.609	0.500	0.355	0.270	0.519	0.500	0.632	0.550	0.742	0.500

and consistency [45]). To estimate SHAP values, we used the `TreeSHAP` algorithm of the `SHAP` Python package (version 0.39.0) on the best-performing ML model for each stress measure, after retraining it on the entire data set of observations. We derived three different types of SHAP plots for all stress measures to gain insights into the workings of our best-performing ML models (i.e., the SHAP global, beeswarm and dependence plots).

5. Results

In this section, we provide a summary of the main results of the ML pipeline from Section 4.6, which lies at the core of the proposed methodology described in Section 4 and shown in Fig. 1. Specifically, in Section 5.1 we describe in detail the results of the modelling procedures and the evaluation of the resulting ML models for all three stress measures and combinations of collected mouse, keyboard and cardiac activity data. Then, in Section 5.2 we provide the results of the SHAP value analysis of the best-performing models emerging from the ML pipeline. To do so, we show (1) global, (2) beeswarm, and (3) dependence SHAP plots [89] for the lowest and highest levels of all three stress measures.

5.1. Evaluation of ML models

In Tables 3 and 4 we provide an overview of the best-performing SVM, RF and LightGBM models, considering all three stress measures. As mentioned in Section 4.6, the best-performing models are chosen by considering the weighted average F1 score over the 10 test data folds. Table 3 contains the results for the models trained on behavioural features, for which observations from block 4 were excluded (i.e., the mouse and/or keyboard modalities). Table 4 contains the results for the models trained on either HRV data only, or on HRV data in combination with one or both of the behavioural modalities, and considering observations drawn from all six experimental blocks. Both tables include the results of the baseline classifiers maximising the class priors. In the Appendix, we indicate the optimal hyperparameter configurations for the best-performing models per stress measure.

Overall, the LightGBM models trained on mouse and keyboard features outperform the models trained on other feature subsets or with other algorithms, for all three stress measures. These models result in F1 scores equal to 0.625, 0.631 and 0.775 for perceived stress, arousal and valence, respectively. All the tuned models in Tables 3 and 4 perform considerably better than the corresponding baseline classifiers. In particular, the LightGBM model trained on mouse and keyboard features to detect perceived stress (F1=0.625) shows an increase in F1 score equal to +33% with respect to the corresponding baseline classifier (F1=0.469). The increase of F1 score in the case of arousal and valence is equal to +72% and +20%, respectively. The performance improvements of the best models compared to the baseline classifiers are driven by gains in both precision and recall. The gains in precision are higher than the ones in recall, since the baseline classifiers optimising class priors naturally achieve a fairly high recall on the imbalanced test folds.

The RF models consistently outperform the SVM models in terms of all four performance measures, while the LightGBM models fare even better for most feature sets and stress measures. The only exception are the HRV modality models, where the tuned RF model achieved a slightly higher F1 score than the LightGBM one (0.531 vs. 0.528). For the SVM algorithm, models trained on mouse features yield slightly better results than on keyboard features, while for RF models, the reverse holds true. In the case of LightGBM, the mouse modality worked better than keyboard data for detecting perceived stress, but not for

Table 5
Average performance results of the best LightGBM on mouse and keyboard features modalities, for each stress measure and corresponding classes.

Label	Class	F1	precision	recall	AUC
Perceived stress	*low*	**0.776**	**0.726**	**0.832**	**0.741**
	medium	0.382	0.450	0.334	0.679
	high	0.386	0.436	0.348	0.761
Arousal	*low*	**0.716**	**0.712**	**0.721**	**0.768**
	medium	0.591	0.580	0.603	0.739
	high	0.343	0.388	0.309	0.760
Valence	*negative*	0.294	0.298	0.303	0.875
	neutral	0.515	0.575	0.468	0.787
	positive	**0.865**	**0.843**	**0.889**	**0.797**

arousal and valence. The combination of the two modalities, however, generally improves performance in comparison to single modalities. For SVM, RF and LightGBM algorithms, models on HRV features are outperformed by those including also one or both behavioural modalities for all stress measures. The increase is generally larger when adding mouse modality rather than the keyboard modality, with the exception of the LightGBM model detecting valence. In the case of the SVM models detecting arousal, the model based on both HRV and keyboard features even performs worse than the model based on HRV data only. Considering the best LightGBM models on mouse and keyboard data in comparison to those based on all features, the addition of HRV features to behavioural ones leads to a decrease of performance equal to −2% for either perceived stress, arousal, or valence. Finally, comparing the three different stress measures, valence is the most accurately detected measure (F1=0.775), followed by arousal and perceived stress, for which the best LightGBM models trained on behavioural measures reach similar performance (F1=0.625 and F1=0.631).

For each stress measure, in Table 5 we report the per class average performance over the 10 imbalanced test folds of the best-performing model, that is the LightGBM model trained on mouse and keyboard features. We recall that, due to class imbalance (see Table 2), each test fold contains on average 140 samples with a *medium* stress level, 61 samples with a *high* arousal level, and only 10 samples with a *negative* valence level. In the case of perceived stress detection, the model reaches the lowest F1 score on the *medium* stress level class. Performance slightly improves in the case of *high* stress class and considerably in the case of the *low* stress class, where F1=0.776. In the case of the detection of arousal and valence, the performance of the model decreases with the increase of class imbalance. In fact, the arousal model shows the highest performance in detecting the *low* arousal level with F1=0.716, followed by *medium* and *high* levels. Similarly, the valence model shows the lowest performance on the *negative* valence class, considerably improving on the *neutral* (F1=0.515) and *positive* classes (F1=0.865), instead.

To verify the beneficial effect the SMOTE oversampling, we checked the performance of the best-performing models (i.e., LightGBM models trained on mouse and keyboard features) when tuned on the same hyperparameter grids (see Appendix) but without SMOTE. Results indicate that without SMOTE, the F1 scores would decrease by 2% for perceived stress and arousal, and 3% for valence.

5.2. Model interpretation with SHAP values

In order to visualise the feature contributions in our best-performing models, we show the SHAP global and beeswarm plots over 20 features for the tuned LightGBM models based on mouse and keyboard data, after retraining on the entire data set (without block 4 observations). Specifically, we include SHAP beeswarm plots for the *low* and *high* perceived stress, *low* and *high* arousal, and the *negative* and *positive* valence classes, respectively. Finally, we present a selection of SHAP dependence plots, for each stress measure.

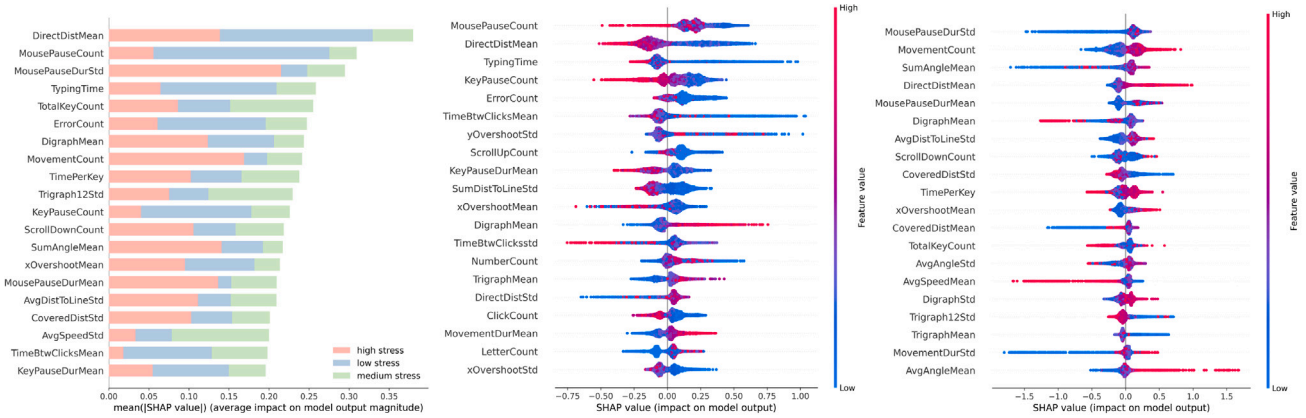
The SHAP global plots include the set of the 20 features with the highest mean absolute SHAP values across all three classes, per each stress measure. The SHAP beeswarm plots list the 20 features with the highest mean absolute SHAP value for the selected class. In each plot, the x-axis corresponds to the SHAP values for all features at each observation, shown as coloured dots. When multiple dots land at the same x position, they are piled up to show density. Their colour corresponds to the values of the feature under consideration, see the vertical scale on the right-hand side of each plot. Note that the features (except key and click counts) have been user-calibrated with data from block 1, and that detailed descriptions of all features can be found in Appendix. By definition, for any feature, observation, and class, a positive SHAP value indicates a positive impact of the corresponding feature on the prediction of the class for the given observation, and vice versa for *negative* values. Finally, each SHAP dependence plot maps the SHAP values of a selected feature against its observed values. The observations can be coloured by the value of another feature, allowing to analyse interactions between pairs of features. The procedure is particularly useful in the case of "vertical dispersion", that is when multiple observations in the dependence plot are characterised by the same feature value but different SHAP values.

5.2.1. Perceived stress

The global SHAP plot in Fig. 4(a) shows that the three features with highest mean absolute SHAP values are all mouse modality features, specifically the mean Euclidean distance between the end and start locations of mouse movements (DirectDistMean), and the count and standard deviation of durations in mouse movement pauses (MousePauseCount, MousePauseDurStd). For DirectDistMean and MousePauseCount, the mean absolute SHAP value for the *low* perceived stress level shows the highest contribution to the total mean absolute SHAP value of the features, while for MousePauseDurStd, the mean absolute SHAP value of the *high* perceived stress level accounts for 73% of the total mean absolute SHAP value of the feature.

The *low* perceived stress level SHAP beeswarm plot (Fig. 4(b)) shows multiple features with a high dispersion of SHAP values around zero. This means that features such as the mean and standard deviation of the duration between clicks (TimeBtwClicksMean, TimeBtwClicksStd), and the standard deviation of the "overshoot" distance in the y direction (yOvershootStd) also contributed strongly to the probability of *low* perceived stress for some observations. Considering the globally important mouse features DirectDistMean and MousePauseCount, we note that low values of either feature are associated with positive SHAP values, for most observations. This means that values of the mean Euclidean distance of mouse movements lower than average (i.e., lower than the corresponding average in block 1, per participant) and a low number of pauses between mouse movements are generally associated with an increase in the probability of *low* perceived stress by the model.

The SHAP dependence plot in Fig. 7(a) plots the SHAP values of DirectDistMean against the feature's values. The colouring by the value of the mean average speed of movements (AvgSpeedMean) shows that the observations with movements of shorter distance than average, which contribute positively to the *low* perceived stress class prediction, are characterised by low average speed. Important keyboard-related features such as the time spent typing (TypingTime), the counts of



(a) Global SHAP plot for perceived stress (b) SHAP swarmplot for *low* perceived stress (c) SHAP swarmplot for *high* perceived stress

Fig. 4. SHAP value plots for the best LightGBM model on mouse and keyboard features, in the case of perceived stress detection. Descriptions of the features can be found in Appendix.

pauses in typing (KeyPauseCount) and delete and backspace key occurrences per minute (ErrorCount) follow a similar overall pattern to the one of DirectDistMean and MousePauseCount, with low values increasing the probability of the *low* perceived stress class.

The *high* perceived stress level SHAP beeswarm plot (Fig. 4(c)) shows that the top three most important features are mouse features, specifically MousePauseDurStd, the number of mouse movements (MovementCount) and the mean of the summed angles traversed within a mouse movement (SumAngleMean). The plot also presents a few mouse features with high dispersion of SHAP values around zero. Considering the globally important mouse features DirectDistMean and MovementCount, we note that high values of the two features are each associated with positive SHAP values, for most observations. This means that observations with many or farther than average movements are generally associated with an increase in the probability of *high* perceived stress. A SHAP dependence plot of DirectDistMean coloured by SumAngleMean (see Fig. 7(b)) reveals that positive SHAP values of the mouse movements with longer than average distances are mostly characterised by high values of the summed traversed movement angles.

5.2.2. Arousal

The global SHAP plot in Fig. 5(a) shows that the top three features with highest mean absolute SHAP values are MousePauseDurStd, KeyPauseCount and duration spent typing per number of keys pressed (TimePerKey). Their per class mean absolute SHAP value distribution differs. In the case of MousePauseDurStd, the mean absolute SHAP value for the *high* arousal level shows the highest contribution to the total mean absolute SHAP value of the feature. In the case of KeyPauseCount and TimePerKey, the highest contributions are driven by the mean absolute SHAP values for the *low* and *medium* levels, respectively.

The SHAP beeswarm plot for the *low* arousal class (Fig. 5(b)) shows that, globally, features characterising the typing pauses, namely KeyPauseCount and the mean and standard deviation of duration (KeyPauseDurMean, KeyPauseDurStd) contribute strongly to the *low* arousal prediction. Here, a low number of typing pauses, and lower mean and variation in typing pause durations than average are each associated with increased probabilities for *low* arousal by the model.

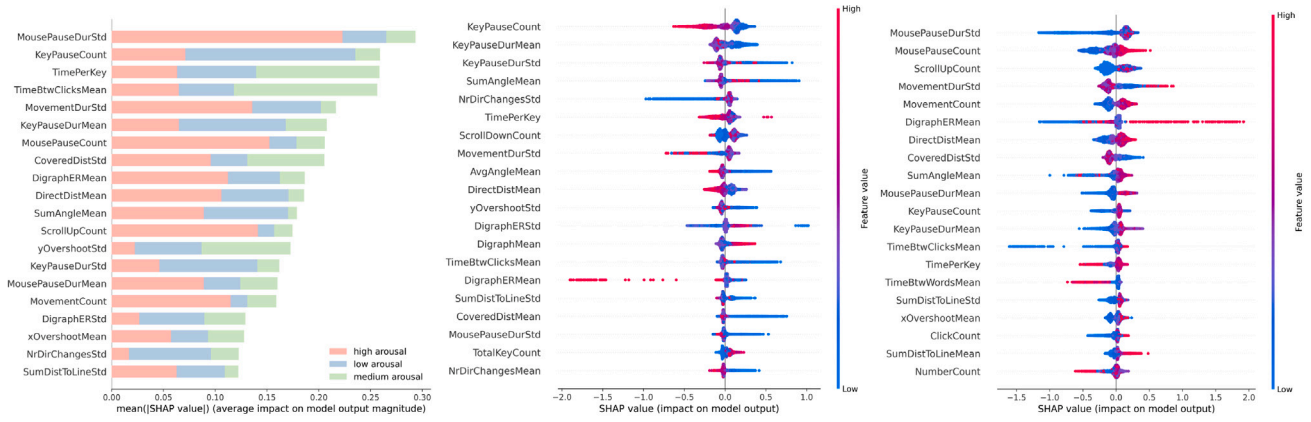
The SHAP beeswarm plot for the *high* arousal class (Fig. 5(c)) shows that the top five features are all mouse behaviour-related. For example, high values of MovementCount and MousePauseCount are each associated with increased probabilities of *high* arousal. The dependence plot in Fig. 7(c) of MovementCount coloured by the value of the MovementDurStd reveals that numbers of mouse movements per minute higher than 26 are generally associated with an increased probability of *high* arousal. Low numbers of mouse movements are associated to a decrease of probability of *high* arousal. Moreover, for each number of mouse movements, higher variability in the duration of the movements is related to a greater reduction in the probability of *high* levels of arousal.

5.2.3. Valence

The global SHAP plot in Fig. 6(a) shows that the top three features with highest mean absolute SHAP values are KeyPauseDurMean, the mean "overshoot" in the x-direction of mouse movements ending in a click (xOvershootMean), and the mean sum of the distances between the recorded positions and the optimal straight line between start and end positions of the mouse movement (SumDistToLineMean). For all three features, the mean absolute SHAP value of the *negative* valence class accounts for the majority of the total mean absolute SHAP value of the feature.

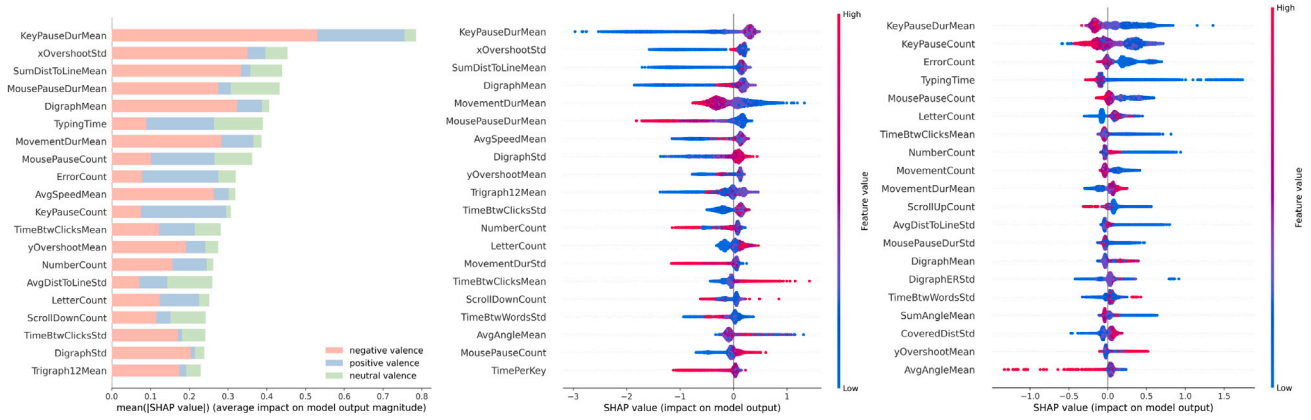
The SHAP beeswarm plot of the *negative* valence class (Fig. 6(b)) shows that globally, low values for KeyPauseDurMean, xOvershootStd, and SumDistToLineMean are each associated with lower probabilities of the *negative* valence class, while low values for MovementDurMean are associated with a higher probability for the *negative* valence class. The first three features show dispersion to the left of the SHAP value zero. Moreover, low values of MovementDurMean are associated with an increased probability of *negative* valence. The dependence plot in Fig. 7(d) of SumDistToLineMean coloured by the value of the mean number of direction changes within a movement (NrDirChangesMean) reveals that mouse movements deviating more strongly from the straight line than average are associated with high probability of *negative* valence and are characterised by a high number of direction changes along the movement. The dependence plot of ErrorCount in Fig. 7(e) shows that a number of typing errors per minute higher than 2 is associated with an increase in the probabilities of *negative* valence.

The SHAP beeswarm plot for the *positive* valence class (Fig. 6(c)) shows that globally, low values for the KeyPauseCount and KeyPauseDurMean, ErrorCount and TypingTime are each associated with an increased probability of *positive* valence by the model. Similarly, a low MousePauseCount is also related to an increased probability of *positive* valence. The dependence plot in Fig. 7(f) of MovementDurMean coloured by the value of the average distance to the optimal straight line between end and start positions of the movement (AvgDistToLineMean) reveals that observations with duration longer than average are associated with an increased probability of *positive* valence, and they are also characterised by high deviations from the optimal straight line. The dependence plot in Fig. 7(g) of KeyPauseCount coloured by the value of ErrorCount reveals that observations with few typing pauses, associated with increased probabilities of *positive* valence, are characterised by a low number of typing errors.



(a) Global SHAP plot for arousal (b) SHAP swarmplot for *low* arousal (c) SHAP swarmplot for *high* arousal

Fig. 5. SHAP value plots for the best LightGBM model on mouse and keyboard features, in the case of arousal detection. Descriptions of the features can be found in Appendix.



(a) Global SHAP plot for valence (b) SHAP swarmplot for *negative* valence (c) SHAP swarmplot for *positive* valence

Fig. 6. SHAP value plots for the best LightGBM model on mouse and keyboard features, in the case of valence detection. Descriptions of the features can be found in Appendix.

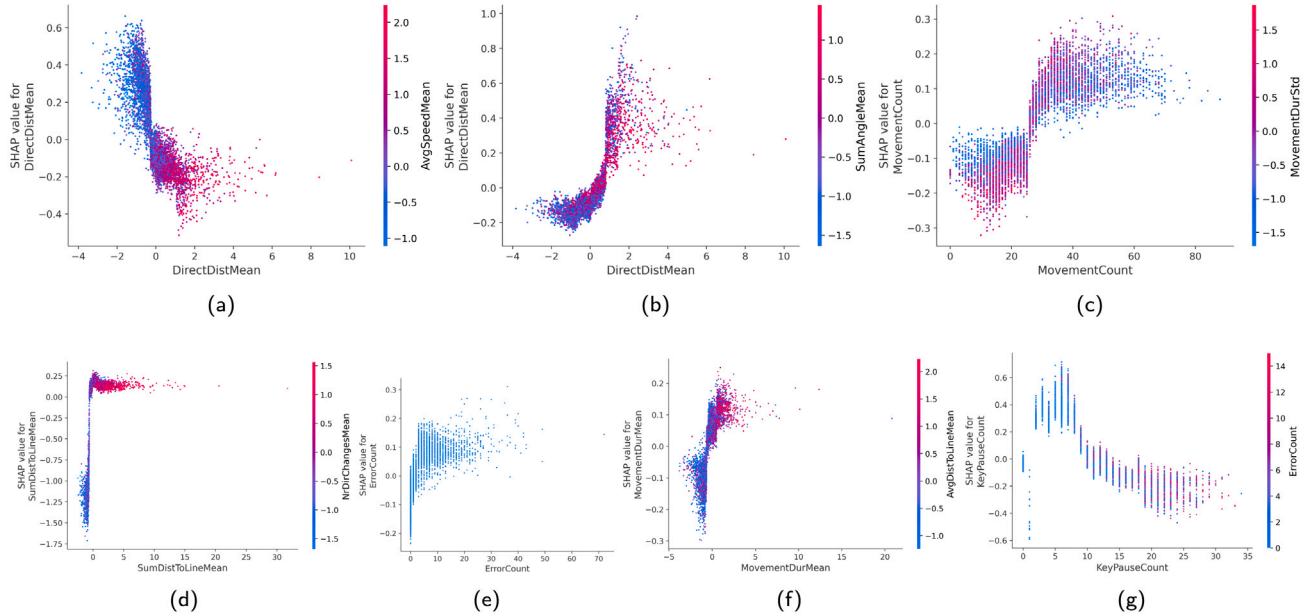


Fig. 7. SHAP dependence plots for the best LightGBM models on mouse and keyboard features for selected features. (a) SHAP values of DirectDistMean coloured by AvgSpeedMean for detecting *low* perceived stress. (b) SHAP values of DirectDistMean coloured by SumAngleMean for detecting *high* perceived stress. (c) SHAP values of MovementCount coloured by MovementDurStd for detecting *high* arousal. (d) SHAP values of SumDistToLineMean coloured by NrDirChangesMean for detecting *negative* valence. (e) SHAP values of ErrorCount for detecting *negative* valence. (f) SHAP values of MovementDurMean coloured by AvgDistToLineMean for detecting *positive* valence. (g) SHAP values of KeyPauseCount coloured by ErrorCount for detecting *positive* valence. Descriptions of the features can be found in Appendix.

Table 6
Overview of key findings and implications.

Our results in context	Our experimental protocol improves the status quo of automated stress detection by simulating a realistic office environment and collecting behavioural and physiological data from a large sample of participants, while the performance of our models is higher or in line with those of comparable studies across the different stress measures.
ML algorithms	LightGBM models outperform SVM and RF in predicting perceived stress, arousal and valence levels.
Data modalities	The combination of mouse and keyboard data outperforms all others in predicting perceived stress, arousal and valence levels.
ML interpretability	SHAP value analyses revealed that the predictions of different levels of stress measures may be characterised by specific mouse movement or typing behaviour: – *high* perceived stress: long, "wavy" mouse movements; *low* perceived stress: short, direct, slow mouse movements, – *high* arousal: high number of mouse movements; *low* arousal: low number of typing pauses, – *positive* valence: low number of typing pauses and typing errors; *negative* valence: high number of typing errors.
Implications for real office environments	Multi-item validated psychological instruments may support the prediction of self-assessed stress more efficiently than ad-hoc scales. A baseline measurement to calibrate the data of new users may help with high inter-individual variability. Research on different time windows and propagation techniques of self-reported stress levels over time is needed. Personalised insights into the predictions using SHAP values may help with finding the right balance of false positives and false negatives. The use of multiple modalities allows managing the time intervals without computer interactions in real office environments. Security concerns regarding the collection of keyboard data in office environments may be mitigated by the derivation of high-level features on local machines.

6. Discussion

In this work, we develop an ML pipeline to detect different stress-related states (perceived stress, arousal, valence) at a granular level (three classes) from physiological (HRV) and behavioural (mouse and keyboard) data sources collected in a laboratory experiment with 90 participants, which simulated a realistic group office environment and induced stress with realistic work-related stressors. The ML pipeline comprises SVM, RF and LightGBM algorithms, the use of SMOTE oversampling to address multiclass imbalance, and the computation of SHAP values to interpret the results of ML modelling. In the following Sections 6.1–6.4 we discuss our key findings from the developed methodology, while in Section 6.5 we consider the implications of our results for future research in real office environments (see Table 6 for a summary of findings and implications).

Our results reveal that LightGBM models trained on mouse and keyboard features were best able to predict perceived stress, arousal and valence levels, in terms of averaged F1 scores on test folds of a stratified 10-fold cross-validation. Highest performance is achieved in the case of valence, followed by arousal and perceived stress, although this difference may be driven largely by the varying degrees of class imbalance (see Table 2). The multiclass labels in our data are moderately to severely imbalanced. This affects the performance of the ML models, especially for the minority classes of all stress measures, as shown in Table 5. However, the models achieve much higher performance on the majority classes of all stress measures. These classes correspond to a state of normal baseline activation for office workers and thus, in a healthy work environment, to the prevalent classes in real life.

6.1. Our results in context

The direct comparison of the results from different stress detection studies, ours included, is impeded by differences in experimental settings, generated features, and the choice of ML algorithms and performance measures. Despite some works achieving very high performances in detecting stress, many authors believe that such results might be overly optimistic due to procedural and methodological limitations of the underlying studies [32,62]. Nevertheless, we try to contextualise our results with a selection of notable works below.

In terms of the experimental protocol used to generate training data, the work by Koldijk et al. [29] is arguably one of the most comparable to our own. Their experiment was also designed to mimic an office scenario, where 25 participants were asked to write reports and prepare presentations under three different conditions: under neutral conditions; under time pressure; and while receiving email messages as a form of interruptions. In contrast, we included an additional socio-evaluative stressor in the form of the adapted TSST-G and tested a larger number of participants. Moreover, we performed a more in-depth analysis of our mouse, keyboard and HRV data resulting in a wider range of features from these modalities. In terms of performance results, Koldijk et al. [29] achieve up to 90% accuracy to distinguish the neutral from stressful working conditions from their entire feature set. However, this set includes facial and posture features derived from camera data, which are likely to raise privacy concerns in real office environments [15]. When considering only computer and physiology data, their achieved accuracy of the binary classification task drops to 68%, which is more in range of our own results. Salmeron-Majadas et al. [63] developed affect recognition models based on mouse and keyboard data generated in a laboratory experiment aimed to simulate a real-life learning scenario. In their experiment, 41 participants were asked to write three short essays based on word prompts of increasing difficulty. Their best models trained on a user-calibrated data set achieved accuracies of 0.647 and 0.663 when classifying three levels of arousal and valence, respectively. As the class-weighted recall is equal to overall accuracy by definition, we may compare those results to the recall values of our own best mouse- and keyboard-based models yielding a similar 0.635 for arousal but considerably higher 0.782 for valence.

Regarding HRV data, Dai et al. [30] recently ran a laboratory experiment where 32 participants were subjected to public speaking, mental arithmetic and cold pressor tasks to elicit stress. HRV data were collected via smartwatches equipped with PPG sensors. Comparable to our own work, Dai et al. [30] employed SVM, RF and boosting methods and further state that they used grid search to tune hyperparameters of their models. When classifying periods perceived as stressful versus non-stressful, the authors achieved an F1 score of 0.599, while we report a value of 0.531 for our best HRV-based model to predict three classes of perceived stress. However, we considerably improve the performance of models trained on HRV data by including behavioural sources, as shown in Tables 3 and 4.

6.2. Machine learning algorithms

Our results show that the boosting-based LightGBM algorithm generally outperforms RF and SVM across feature sets and stress measures, while RF consistently fared better than SVM. There are still relatively few stress detection works implementing boosting algorithms, and those that do typically focus on older methods, such as adaptive

boosting [30,67,69]. In those works, however, boosting algorithms do not always outperform RF [69], and SVM has even beaten both RF and boosting methods on some occasions [30,67]. For example, Can et al. [71] found that RF performed better than SVM on larger feature sets. We also confirmed that applying SMOTE in order to address the imbalance in our data can help increase performance, in line with results from other works [63,65]. With regards to our hyperparameter tuning, it is not possible to compare our results with existing studies, as these only rarely mention any hyperparameter selection strategy [30,62], and even less often report a hyperparameter grid [75]. We hope that by doing so ourselves, we may help with the current lack of comparability of ML-based stress detection models [72].

6.3. Data modalities

In the literature, the link between the physiological stress response and changes in HRV is well-established [for a review, cf. 51], and studies using HRV data to infer relaxed versus stressful states have often achieved high performance results [33,54,66]. In contrast, studies with HRV data where the non-stressed phase corresponds to a basic workload task rather than full relaxation usually see lower performances [30,67]. In our study, the single modality HRV models also attain low performance scores, getting outperformed by both mouse and keyboard single modality models—although they still performed better than the baseline classifier maximising class priors.

The combination of different modalities is generally thought to increase performance in stress detection tasks [22]. Similarly, we found that models using both mouse and keyboard features have higher performance than those trained on only a single modality. Interestingly, when comparing the results for models trained on mouse and/or keyboard features to those additionally including HRV features, it seems that the inclusion of the HRV modality does not yield a large impact on performance. While few studies with both behavioural and physiological features exist [29,34,70], the latter are thought to have more predictive power for stress detection [15]. In contrast, our results add to the existing evidence that the computer-based features may outperform physiological features in the context of work-related stress [29,70,90]. Here, it is possible that chosen stressors elicited a milder physiological stress response than other acute stress tests, or that the baseline workload tasks already incurred a certain level of activation of the physiological system.

Compared to HRV data, few previous studies have explored the use of mouse and keyboard data to detect stress [15,22]. While these two modalities are frequently studied in combination [e.g., 29,63], they are often collected in two separate data generation tasks [32,61]. For example, Pepa et al. [61] developed a 3-level stress detection system based on mouse and keyboard features. To generate keyboard data, participants copied a short text, with and without time and performance pressure, while mouse data were collected in three separate game-like tasks of varying difficulty (e.g., a Tower of Hanoi game). The two disjoint data sets were then used to train two separate classifiers. Such a dichotomous approach might neglect the potential of the combination of both modalities to improve stress detection performance as evidenced by our own results. Our results also indicate that mouse features may have more predictive power to classify stress-related states than keyboard features, especially when used in combination with HRV data. An explanation for this finding might be that stress-induced neuromotor noise could more easily be detected in mouse movements than in typing behaviour. In fact, Freihaut and Göritz [62] have recently reviewed the theoretical reasoning and empirical evidence for the link between stress and mouse movement behaviour. They state that there is strong support for stress affecting goal-directed movements such as mouse movements [62]. In their own experiment with 53 participants, however, the authors found that their trained classification models cannot significantly outperform random guessing in most cases. Hence further research is needed regarding the link between stress and keyboard dynamics, despite promising results from empirical studies [32,59–61], and now this work.

6.4. Interpretation of results with SHAP values

The SHAP value analysis on the best-performing models indicates that both mouse and keyboard features show high importance, for all stress measures, but that the feature attributions differ for each stress measure and across the respective levels. A number of insights emerge from the analysis of the model SHAP beeswarm plots. For example, *low* perceived stress may be characterised by relatively short, slow and direct mouse movements, while those during *high* perceived stress are longer and "wavy" (i.e., they covered a large sum of absolute angles along each movement). A relatively high number of mouse movements may be indicative of *high* levels of arousal, especially if the variability in movement duration is high. *Positive* valence movements could be characterised as relatively long in duration, and deviating far from the straight line, while *negative* valence movements seem to be less precise and show a high number of direction changes. In addition, few typing pauses with a low number of typing errors were indicative of *positive* valence, while a high number of typing errors increased the probability of *negative* valence.

We note that these scenarios stem from a non-exhaustive exploratory analysis and thus need to be confirmed in further studies. We also do not know how well these scenarios might generalise to the real world, where there is likely a high amount of variability in keyboard and mouse usage pattern between companies, employees and even tasks. However, we specifically designed our baseline workload tasks to reflect typical mouse and keyboard usage and selected mouse and keyboard features which characterise general behaviour rather task-specific pattern. To the best of our knowledge, we are the first to investigate different SHAP value plots of mouse and keyboard features in stress detection models. However, a few previous studies have explored statistical analyses in term of correlations or group mean differences of individual mouse or keyboard features with respect to stress levels. For example, and similar to our interpretations of mouse movements relating to perceived stress, Freihaut and Göritz [62] found that participants travelled further distances with their cursors during the *high* stress condition. Sanchez et al. [91] also report a positive and significant correlation of traversed pixels with stress level in a longitudinal field study and they found a significant correlation of the number of errors with stress level. Banholzer et al. [92] found evidence for a trade-off between mouse speed and accuracy under stress. Furthermore, Salmeron-Majadas et al. [58] found a high correlation between valence and the mean duration of mouse movements, in accordance with our findings.

We argue that the value of using interpretable ML methods, such as SHAP, is twofold. First, they become a tool for engineers to audit and validate their ML models, providing valuable insights to understand stress as a function of behavioural and physiological data. Second, as commonly stated in the iML research domain [39,40], they may become valuable in supporting users' understanding of the logic behind the detection of stress levels. In the context of a stress management JITAI, improved transparency of predictions could foster the user's trust in a model, and thereby increase their adherence to the programme [39]. Here, engineers may use local and global SHAP explanations of the ML models and their predictions to provide explanatory reports to the user. For example, Jaber et al. [41] recently designed a stress prediction report aimed at psychiatrists, which is reminiscent of a standard blood test and lists the SHAP value-based prediction impacts of a range of physiological features from the wearable stress and affect detection (WESAD) data set [69]. In addition, the insights from SHAP analyses could potentially also be used to tailor notifications or intervention prompts in a JITAI, supporting the promotion of individualised coping strategies.

6.5. Implications for real office environments

Stress detection in real-life office environments clearly faces many additional challenges over controlled laboratory settings. In the field, ground truth labels are usually collected via self-reported EMA since, unlike in laboratory experiments, there is normally no objective information available on the presence of stressors. As shown by the results of this study, stress can be measured with different measures and ML models show different levels of performance, when considering perceived stress, arousal or valence, hence assessing stress along only one dimension might be inadequate for in-field applications. For example, acute stress is sometimes defined as a state of high arousal combined with negative valence [92]. Our results show that multi-item validated psychological instruments may support the prediction of self-assessed measures of stress more efficiently than ad-hoc scales, such as the one we introduced to measure perceived stress levels. Self-reports may be well-suited to capture inter-individual differences in the response to stressors, but they come with the drawback of relying on peoples' recall abilities, their degree of self-awareness [22], and their willingness to share their genuinely perceived state of stress [47], which further underlines the need for continuous, unobtrusive, data-driven stress detection.

The high inter-individual variability in the stress response means that stress detection models do not generalise well to data from unseen subjects [29,63,82], which is arguably one of the biggest challenges researchers are facing when moving from constrained laboratory environments with limited amounts of participants to deployment in a real office with an abundance of potential new users [23,28]. While a general one-fits-all model is unlikely to produce satisfactory results in the wild, training a new, individual model for each new user is presumably too inefficient. More promising avenues are the use of a baseline measurement either to calibrate the data of new users (which would be a straight-forward adaption for our developed pipeline), or to apply transfer learning on a pretrained model [80].

Moreover, the time window of validity of each self-report has to be assessed. In this work, we backpropagated the self-assessed stress levels over each block (i.e., between 10 and 18 min) for each participant. However, in the field researchers would need to define the time window for which each EMA self-report of stress levels should hold. This time window might be a couple of minutes to an hour or more into the past, or the near future, for each participant. To the best of our knowledge, there is no research on the theoretical foundation for the temporal validity of EMA self-reported stress levels and the optimal frequency of their collection in office environments.

It is also not a priori clear what level of performance an ML model for stress detection in an office environment should achieve. For example, while false negatives increase the risk of missing to adequately treat a person at risk for serious stress-induced health consequences, too many false positives could simply impede a user's workflow, increasing levels of stress and the likelihood of dropout from the stress management programme. Here, the provision of personalised insights into the predictions using SHAP values, and the scheduling of personalised ML model retraining procedures are possible avenues to improve the performance of stress detection in the long run. In addition to iML-based explanations, the provision of context information (e.g., location or current activity) could also support potential users with evaluating and understanding stress-level predictions [93].

Finally, in office employees might not be constantly interacting with a computer throughout their workday. Therefore, when designing ML models detecting stress levels it may be necessary to consider the use of multiple feature modalities, and introduce an appropriate logic for the management of time intervals without computer interactions. We propose the combination of HRV, mouse and keyboard data for this purpose, which we believe to be well-suited for stress detection in office environments in terms of both obtrusiveness and safe-guarding of privacy. Lastly, we note that typing recordings may pose a certain security or privacy risk in a real-world work environment. However, this issue might easily be overcome by relying only on features which do not allow inference of the typed content—such as the ones we have used in this work, and by generating any features directly on local machines.

6.6. Limitations of our work

This work used data generated in a single session experiment with a simulated office environment, artificial stressors and participants that were healthy and predominantly young, Caucasian University students. Clearly, in a laboratory experiment we could only approximate real-life conditions up to a certain degree, especially as participants were asked to follow a hypothetical scenario which might differ significantly from their own reality and which lasted only a couple of hours. While the experimental software and tasks were designed to be reflective of typical menial office clerk work—engaging but straight-forward, participants did not have any prior experience with the experiment software before the study. In a real office, employees naturally complete a much wider array of tasks of varying complexities, and they might have years of experience with them. The Firstbeat Bodyguard 2 devices used in this study are two-electrode ECG devices allowing for high quality collection of cardiac activity data. However, electrode-based devices are too cumbersome and obtrusive to be worn continuously during working hours. Although these devices might not offer the same level of data quality, future field studies should consider using less obtrusive PPG devices that can be worn on the arm or wrist to collect physiological data. While the data set in our study is larger than those used in many comparable studies, 6776 observations are still relatively few for ML modelling purposes with data-intensive methods. Finally, although we derived different behavioural features based on previous literature, we were limited by the software used to record mouse and keyboard inputs.

7. Conclusion

In this work, we propose an ML-based methodology for the automated detection of a stress measures that relies on mouse, keyboard and HRV data. The data were generated in a laboratory experiment reflecting a realistic group office environment with 90 participants, where both the baseline workload and the stress eliciting procedures were designed to emulate real office situations. We find that a Light gradient boosting algorithm achieves the best performance for the detection of perceived stress, arousal and valence states from a combination of mouse and keyboard features. The use of iML methods, namely SHAP values, reveals a number of behavioural patterns associated to *low* and *high* perceived stress, *low* and *high* arousal, and *negative* and *positive* valence levels. Our study fills different methodological gaps in the research on the automated detection of stress with ML methods, namely the closer approximation of real office conditions in a laboratory experiment, the conjoint use of mouse and keyboard features as well as the combination of both physiological and behavioural data modalities, advanced boosting algorithms and extensive hyperparameter tuning, and the interpretation of resulting models with SHAP values. In the future, field studies are needed to confirm the results from the simulated laboratory environment at scale in real-life conditions with actual employees, in order to further advance the research on automated, personalised, continuous detection of stress levels in office environments and to ultimately enable the development of JITAI programmes for stress management at the workplace.

CRediT authorship contribution statement

Mara Naegelin: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – original draft, Visualization, Project administration. **Raphael P. Weibel:** Conceptualization, Methodology, Software, Investigation, Data Curation, Writing – original draft, Visualization, Project administration. **Jasmine I. Kerr:** Conceptualization, Methodology, Investigation, Data Curation, Writing – original draft, Visualization, Project administration. **Victor R. Schinazi:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Roberto La Marca:** Conceptualization, Resources, Writing – review & editing, Funding acquisition. **Florian von Wangenheim:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Funding acquisition. **Christoph Hoelscher:** Writing – review & editing, Funding acquisition. **Andrea Ferrario:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study is part of a larger project and supported by the Donald C. Cooper Fonds via the ETH Zurich (PS: 1-004675-000). We thank Dr. Stefan Wehrli and Giordano Giannoccolo from ETH Zurich's Decision Science Laboratory for their assistance and Erika Meins, Sebastian Tillmanns, Marcus Zimmer, Anita Schärer, Hantao Zhao, Amray Schwabe, Caterina Bérubé, Yanick Lukic and Sabrina Trachsler for their supporting roles as research assistants and actors.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104299.

References

[1] K. Sparks, B. Faragher, C.L. Cooper, Well-being and occupational health in the 21st century workplace, J. Occup. Organ. Psychol. 74 (4) (2001) 489–509, http://dx.doi.org/10.1348/096317901167497.

[2] A. Soylu, S.S. Campbell, Physical and emotional stresses of technology on employees in the workplace, J. Employ. Couns. 49 (3) (2012) 130–139, http://dx.doi.org/10.1002/j.2161-1920.2012.00013.x.

[3] T. Cox, A. Griffiths, E. Rial-González, Research on Work-related Stress, Technical Report, European Agency for Safety and Health at Work, Luxembourg, 2000.

[4] H. Russell, B. Maître, D. Watson, É. Fahey, Job stress and working conditions: Ireland in comparative perspective, Technical Report, (RS84) The Economic and Social Research Institute (ESRI), Dublin, Ireland, 2018, http://dx.doi.org/10.26504/rs84.

[5] R. Krieger, M. Arial, Arbeitsbedingungen und Gesundheit: Stress [Working conditions and health: Stress], Technical Report, Staatssekretariat für Wirtschaft (SECO), Bern, Switzerland, 2020.

[6] C. Goetz, R. Bavaresco, R. Kunst, J. Barbosa, Industrial intelligence in the care of workers' mental health: A review of status and challenges, Int. J. Ind. Ergon. 87 (2022) 103234, http://dx.doi.org/10.1016/j.ergon.2021.103234.

[7] P. Majumdar, A. Biswas, S. Sahu, COVID-19 pandemic and lockdown: cause of sleep disruption, depression, somatic pain, and increased screen exposure of office workers and students of India, Chronobiol. Int. 37 (8) (2020) 1191–1200, http://dx.doi.org/10.1080/07420528.2020.1786107.

[8] D. de Quervain, A. Aerni, E. Amini, D. Bentz, D. Coynel, C. Gerhards, V. Freytag, A. Papassotiropoulos, N. Schicktanz, T. Schlitt, A. Zimmer, P. Zuber, The Swiss Corona Stress Study: November 2021, 2021, OSF Preprints. http://dx.doi.org/10.31219/osf.io/x6zu7.

[9] G.P. Chrousos, Stress and disorders of the stress system, Nat. Rev. Endocrinol. 5 (7) (2009) 374–381, http://dx.doi.org/10.1038/nrendo.2009.106.

[10] S. Stansfeld, B. Candy, Psychosocial work environment and mental health–a meta-analytic review, Scand. J. Work, Environ. Health 32 (6) (2006) 443–462, http://dx.doi.org/10.5271/sjweh.1050.

[11] M. Kivimäki, I. Kawachi, Work stress as a risk factor for cardiovascular disease, Curr. Cardiol. Rep. 17 (9) (2015) http://dx.doi.org/10.1007/s11886-015-0630-8.

[12] K.R. Lorig, H.R. Holman, Self-management education: History, definition, outcomes, and mechanisms, Ann. Behav. Med. 26 (1) (2003) 1–7, http://dx.doi.org/10.1207/S15324796ABM2601_01.

[13] S. Gonul, T. Namli, S. Huisman, G.B. Laleci Erturkmen, I.H. Toroslu, A. Cosar, An expandable approach for design and personalization of digital, just-in-time adaptive interventions, J. Am. Med. Inform. Assoc. 26 (3) (2019) 198–210, http://dx.doi.org/10.1093/jamia/ocy160.

[14] A. Rozet, I.M. Kronish, J.E. Schwartz, K.W. Davidson, Using machine learning to derive just-in-time and personalized predictors of stress: Observational study bridging the gap between nomothetic and ideographic approaches, J. Med. Internet Res. 21 (4) (2019) e12910, http://dx.doi.org/10.2196/12910.

[15] D. Carneiro, P. Novais, J.C. Augusto, N. Payne, New methods for stress assessment and monitoring at the workplace, IEEE Trans. Affect. Comput. 10 (2) (2019) 237–254, http://dx.doi.org/10.1109/taffc.2017.2699633.

[16] K.E. Heron, J.M. Smyth, Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments, Br. J. Health Psychol. 15 (1) (2010) 1–39, http://dx.doi.org/10.1348/135910709X466063.

[17] I. Nahum-Shani, S.N. Smith, B.J. Spring, L.M. Collins, K. Witkiewitz, A. Tewari, S.A. Murphy, Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support, Ann. Behav. Med. 52 (6) (2018) 446–462, http://dx.doi.org/10.1007/s12160-016-9830-8.

[18] F. Künzler, J.-N. Kramer, T. Kowatsch, Efficacy of mobile context-aware notification management systems: A systematic literature review and meta-analysis, in: Proc. 13th IEEE Int. Conf. Wireless Mob. Comput., Netw. Commun., WiMob, 2017, pp. 131–138, http://dx.doi.org/10.1109/wimob.2017.8115839.

[19] R.S. Sadasivam, E.M. Borglund, R. Adams, B.M. Marlin, T.K. Houston, Impact of a collective intelligence tailored messaging system on smoking cessation: The PERSPeCT randomized experiment, J. Med. Internet Res. 18 (11) (2016) e285, http://dx.doi.org/10.2196/jmir.6465.

[20] L. Wang, L.C. Miller, Just-in-the-moment adaptive interventions (JITAI): A meta-analytical review, Health Commun. 35 (12) (2020) 1531–1544, http://dx.doi.org/10.1080/10410236.2019.1652388.

[21] L.P.S. Dias, J.L.V. Barbosa, L.P. Feijó, H.D. Vianna, Development and testing of iAware model for ubiquitous care of patients with symptoms of stress, anxiety and depression, Comput. Methods Programs Biomed. 187 (2020) 105113, http://dx.doi.org/10.1016/j.cmpb.2019.105113.

[22] A. Alberdi, A. Aztiria, A. Basarab, Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review, J. Biomed. Inform. 59 (2016) 49–75, http://dx.doi.org/10.1016/j.jbi.2015.11.007.

[23] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, M. Tsiknakis, Review on psychological stress detection using biosignals, IEEE Trans. Affect. Comput. 13 (1) (2022) 440–460, http://dx.doi.org/10.1109/taffc.2019.2927337.

[24] S. Gradl, M. Wirth, T. Zillig, B.M. Eskofier, Visualization of heart activity in virtual reality: A biofeedback application using wearable sensors, in: Proc. 15th IEEE Int. Conf. Wearable Implantable Body Sens. Netw., BSN, 2018, pp. 152–155, http://dx.doi.org/10.1109/bsn.2018.8329681.

[25] R. Bavaresco, H. Arruda, E. Rocha, J. Barbosa, G.-P. Li, Internet of Things and occupational well-being in industry 4.0: A systematic mapping study and taxonomy, Comput. Ind. Eng. 161 (2021) 107670, http://dx.doi.org/10.1016/j.cie.2021.107670.

[26] S. Gedam, S. Paul, A review on mental stress detection using wearable sensors and machine learning techniques, IEEE Access 9 (2021) 84045–84066, http://dx.doi.org/10.1109/access.2021.3085502.

[27] Y.S. Can, B. Arnrich, C. Ersoy, Stress detection in daily life scenarios using smart phones and wearable sensors: A survey, J. Biomed. Inform. 92 (2019) 103139, http://dx.doi.org/10.1016/j.jbi.2019.103139.

[28] E. Smets, W. De Raedt, C. Van Hoof, Into the wild: The challenges of physiological stress detection in laboratory and ambulatory settings, IEEE J. Biomed. Health Inform. 23 (2) (2019) 463–473, http://dx.doi.org/10.1109/jbhi.2018.2883751.

[29] S. Koldijk, M.A. Neerincx, W. Kraaij, Detecting work stress in offices by combining unobtrusive sensors, IEEE Trans. Affect. Comput. 9 (2) (2018) 227–239, http://dx.doi.org/10.1109/taffc.2016.2610975.

[30] R. Dai, C. Lu, L. Yun, E. Lenze, M. Avidan, T. Kannampallil, Comparing stress prediction models using smartwatch physiological signals and participant self-reports, Comput. Methods Programs Biomed. 208 (2021) 106207, http://dx.doi.org/10.1016/j.cmpb.2021.106207.

[31] S.S. Panicker, P. Gayathri, A survey of machine learning techniques in physiology based mental stress detection systems, Biocybern. Biomed. Eng. 39 (2) (2019) 444–469, http://dx.doi.org/10.1016/j.bbe.2019.01.004.

[32] D.R. da Cunha Silva, Z. Wang, R. Gutierrez-Osuna, Towards participant-independent stress detection using instrumented peripherals, IEEE Trans. Affect. Comput. (2021) http://dx.doi.org/10.1109/taffc.2021.3061417.

[33] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang, Z. Zhao, Detecting work-related stress with a wearable device, Comput. Ind. 90 (2017) 42–49, http://dx.doi.org/10.1016/j.compind.2017.05.004.

[34] K. Yang, C. Wang, Y. Gu, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, J. Goncalves, Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition, IEEE Trans. Affect. Comput. (2021) http://dx.doi.org/10.1109/taffc.2021.3100868.

[35] L.G. Jaimes, M. Llofriu, A. Raij, PREVENTER, a selection mechanism for just-in-time preventive interventions, IEEE Trans. Affect. Comput. 7 (3) (2016) 243–257, http://dx.doi.org/10.1109/taffc.2015.2490062.

[36] A. Sano, P. Johns, M. Czerwinski, Designing opportune stress intervention delivery timing using multi-modal data, in: Proc. 7th Int. Conf. Affective Comput. Intell. Interact., ACII, 2017, pp. 346–353, http://dx.doi.org/10.1109/acii.2017.8273623.

[37] D. Peters, R.A. Calvo, R.M. Ryan, Designing for motivation, engagement and wellbeing in digital experience, Front. Psychol. 9 (2018) 797, http://dx.doi.org/10.3389/fpsyg.2018.00797.

[38] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inform. 113 (2021) 103655, http://dx.doi.org/10.1016/j.jbi.2020.103655.

[39] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, http://dx.doi.org/10.1016/j.artint.2018.07.007.

[40] A. Ferrario, M. Loi, How explainability contributes to trust in AI, in: Proc. 2022 ACM Conf. Fairness Account. Transpar., FAccT, 2022, pp. 1457–1466, http://dx.doi.org/10.1145/3531146.3533202.

[41] D. Jaber, H. Hajj, F. Maalouf, W. El-Hajj, Medically-oriented design for explainable AI for stress prediction from physiological measurements, BMC Med. Inform. Decis. Mak. 22 (1) (2022) 38, http://dx.doi.org/10.1186/s12911-022-01772-2.

[42] B. von Dawans, C. Kirschbaum, M. Heinrichs, The Trier social stress test for groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format, Psychoneuroendocrinology 36 (4) (2011) 514–522, http://dx.doi.org/10.1016/j.psyneuen.2010.08.004.

[43] J.I. Kerr, M. Naegelin, R.P. Weibel, A. Ferrario, R. La Marca, F. von Wangenheim, C. Hoelscher, V.R. Schinazi, The effects of acute work stress and appraisal on psychobiological stress responses in a group office environment, Psychoneuroendocrinology 121 (2020) 104837, http://dx.doi.org/10.1016/j.psyneuen.2020.104837.

[44] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357, http://dx.doi.org/10.1613/jair.953.

[45] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67, http://dx.doi.org/10.1038/s42256-019-0138-9.

[46] R.S. Lazarus, S. Folkman, Stress, Appraisal, and Coping, Springer-Verlag, New York, 1984.

[47] E.S. Epel, A.D. Crosswell, S.E. Mayer, A.A. Prater, G.M. Slavich, E. Puterman, W.B. Mendes, More than a feeling: A unified view of stress measurement for population science, Front. Neuroendocrinol. 49 (2018) 146–169, http://dx.doi.org/10.1016/j.yfrne.2018.03.001.

[48] A. Iacovides, K. Fountoulakis, S. Kaprinis, G. Kaprinis, The relationship between job stress, burnout and clinical depression, J. Affect. Disord. 75 (3) (2003) 209–221, http://dx.doi.org/10.1016/S0165-0327(02)00101-5.

[49] M. Sharmin, A. Raij, D. Epstien, I. Nahum-Shani, J.G. Beck, S. Vhaduri, K. Preston, S. Kumar, Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions, in: Proc. 2015 ACM Int. Joint Conf. Pervasive Ubiquitous Comput., UbiComp, 2015, pp. 505–516, http://dx.doi.org/10.1145/2750858.2807537.

[50] J. Campbell, U. Ehlert, Acute psychosocial stress: Does the emotional stress response correspond with physiological responses? Psychoneuroendocrinology 37 (8) (2012) 1111–1134, http://dx.doi.org/10.1016/j.psyneuen.2011.12.010.

[51] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y.H. Lee, B.-H. Koo, Stress and heart rate variability: A meta-analysis and review of the literature, Psychiatry Investig. 15 (3) (2018) 235–245, http://dx.doi.org/10.30773/pi.2017.08.17.

[52] F. Shaffer, R. McCraty, C.L. Zerr, A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability, Front. Psychol. 5 (2014) 1040, http://dx.doi.org/10.3389/fpsyg.2014.01040.

[53] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, L. Pecchia, Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis, Biomed. Signal Process. Control 18 (2015) 370–377, http://dx.doi.org/10.1016/j.bspc.2015.02.012.

[54] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, S. Kumar, cStress: Towards a gold standard for continuous stress assessment in the mobile environment, in: Proc. 2015 ACM Int. Joint Conf. Pervasive Ubiquitous Comput., UbiComp, 2015, pp. 493–504, http://dx.doi.org/10.1145/2750858.2807526.

[55] J. Kallio, E. Vildjiounaite, J. Kantorovitch, A. Kinnula, M.B. López, Unobtrusive continuous stress detection in knowledge work—Statistical analysis on user acceptance, Sustainability 13 (4) (2021) 2003, http://dx.doi.org/10.3390/su13042003.

[56] A.W.A. Van Gemmert, G.P. Van Galen, Stress, neuromotor noise, and human performance: A theoretical perspective, J. Exp. Psychol.: Hum. Percept. Perform. 23 (5) (1997) 1299–1313, http://dx.doi.org/10.1037//0096-1523.23.5.1299.

[57] T. Yamauchi, Mouse trajectories and state anxiety: Feature selection with random forest, in: Proc. 2013 Humaine Assoc. Conf. Affective Comput. Intell. Interact., ACII, 2013, pp. 399–404, http://dx.doi.org/10.1109/acii.2013.72.

[58] S. Salmeron-Majadas, O.C. Santos, J.G. Boticario, An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context, in: Proc. 18th Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst., KES, in: Procedia Comput. Sci., 35, 2014, pp. 691–700, http://dx.doi.org/10.1016/j.procs.2014.08.151.

[59] L.M. Vizer, L. Zhou, A. Sears, Automated stress detection using keystroke and linguistic features: An exploratory study, Int. J. Hum.-Comput. Stud. 67 (10) (2009) 870–886, http://dx.doi.org/10.1016/j.ijhcs.2009.07.005.

[60] C. Epp, M. Lippold, R.L. Mandryk, Identifying emotional states using keystroke dynamics, in: Proc. SIGCHI Conf. Hum. Factors Comput. Syst., CHI, 2011, pp. 715–724, http://dx.doi.org/10.1145/1978942.1979046.

[61] L. Pepa, A. Sabatelli, L. Ciabattoni, A. Monteriù, F. Lamberti, L. Morra, Stress detection in computer users from keyboard and mouse dynamics, IEEE Trans. Consum. Electron. 67 (1) (2021) 12–19, http://dx.doi.org/10.1109/tce.2020.3045228.

[62] P. Freihaut, A.S. Göritz, Using the computer mouse for stress measurement – An empirical investigation and critical review, Int. J. Hum.-Comput. Stud. 145 (2021) 102520, http://dx.doi.org/10.1016/j.ijhcs.2020.102520.

[63] S. Salmeron-Majadas, R.S. Baker, O.C. Santos, J.G. Boticario, A machine learning approach to leverage individual keyboard and mouse interaction behavior from multiple users in real-world learning scenarios, IEEE Access 6 (2018) 39154–39179, http://dx.doi.org/10.1109/access.2018.2854966.

[64] J. Aigrain, M. Spodenkiewicz, S. Dubuiss, M. Detyniecki, D. Cohen, M. Chetouani, Multimodal stress detection from multiple assessments, IEEE Trans. Affect. Comput. 9 (4) (2018) 491–506, http://dx.doi.org/10.1109/taffc.2016.2631594.

[65] A.S. Anusha, J. Jose, S.P. Preejith, J. Jayaraj, S. Mohanasankar, Physiological signal based work stress detection using unobtrusive sensors, Biomed. Phys. Eng. Express 4 (6) (2018) 065001, http://dx.doi.org/10.1088/2057-1976/aadbd4.

[66] S. Betti, R. Molino Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. Del Ry, 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. 65 (8) (2017) 1748–1758, http://dx.doi.org/10.1109/tbme.2017.2764507.

[67] M. Gjoreski, M. Luštrek, M. Gams, H. Gjoreski, Monitoring stress with a wrist device using context, J. Biomed. Inform. 73 (2017) 159–170, http://dx.doi.org/10.1016/j.jbi.2017.08.006.

[68] F. Di Martino, F. Delmastro, High-resolution physiological stress prediction models based on ensemble learning and recurrent neural networks, in: Proc. 2020 IEEE Symp. Comput. Commun., ISCC, 2020, pp. 1–6, http://dx.doi.org/10.1109/iscc50000.2020.9219716.

[69] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing WESAD, a multimodal dataset for wearable stress and affect detection, in: Proc. 20th ACM Int. Conf. Multimodal Interact., ICMI, 2018, pp. 400–408, http://dx.doi.org/10.1145/3242969.3242985.

[70] H. Zhang, L. Cao, L. Feng, M. Yang, Multi-modal interactive fusion method for detecting teenagers' psychological stress, J. Biomed. Inform. 106 (2020) 103427, http://dx.doi.org/10.1016/j.jbi.2020.103427.

[71] Y.S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, C. Ersoy, Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches, IEEE Access 8 (2020) 38146–38163, http://dx.doi.org/10.1109/access.2020.2975351.

[72] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen, D. Kotz, Evaluating the reproducibility of physiological stress detection models, in: Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., IMWUT, 4, (4) 2020, 147, http://dx.doi.org/10.1145/3432220.

[73] V. Bremer, P.I. Chow, B. Funk, F.P. Thorndike, L.M. Ritterband, Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: Machine learning approach, J. Med. Internet Res. 22 (10) (2020) e17738, http://dx.doi.org/10.2196/17738.

[74] M. Sultana, M. Al-Jefri, J. Lee, Using machine learning and smartphone and smartwatch data to detect emotional states and transitions: Exploratory study, JMIR MHealth UHealth 8 (9) (2020) e17818, http://dx.doi.org/10.2196/17818.

[75] S. Huang, J. Li, P. Zhang, W. Zhang, Detection of mental fatigue state with wearable ECG devices, Int. J. Med. Inform. 119 (2018) 39–46, http://dx.doi.org/10.1016/j.ijmedinf.2018.08.010.

[76] A. Ng, B. Wei, J. Jain, E.A. Ward, S.D. Tandon, J.T. Moskowitz, S. Krogh-Jespersen, L.S. Wakschlag, N. Alshurafa, Predicting the next-day perceived and physiological stress of pregnant women by using machine learning and explainability: Algorithm development and validation, JMIR MHealth UHealth 10 (8) (2022) e33850, http://dx.doi.org/10.2196/33850.

[77] R. Steyer, P. Schwenkmezger, P. Notz, M. Eid, Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF) [Multidimensional Mood Questionnaire], Hogrefe, Göttingen, Germany, 1997.

[78] R. Castaldo, L. Montesinos, P. Melillo, C. James, L. Pecchia, Ultra-short term HRV features as surrogates of short term HRV: a case study on mental stress detection in real life, BMC Med. Inform. Decis. Mak. 19 (2019) 12, http://dx.doi.org/10.1186/s12911-019-0742-y.

[79] T. Pereira, P.R. Almeida, J.P.S. Cunha, A. Aguiar, Heart rate variability metrics for fine-grained stress level assessment, Comput. Methods Programs Biomed. 148 (2017) 71–80, http://dx.doi.org/10.1016/j.cmpb.2017.06.018.

[80] A. Maxhuni, P. Hernandez-Leal, L.E. Sucar, V. Osmani, E.F. Morales, O. Mayora, Stress modelling and prediction in presence of scarce data, J. Biomed. Inform. 63 (2016) 344–356, http://dx.doi.org/10.1016/j.jbi.2016.08.023.

[81] R. Wampfler, S. Klingler, B. Solenthaler, V.R. Schinazi, M. Gross, C. Holz, Affective state prediction from smartphone touch and sensor data in the wild, in: Proc. 2022 CHI Conf. Hum. Factors Comput. Syst., CHI, 2022, 403, http://dx.doi.org/10.1145/3491102.3501835.

[82] D. Carneiro, P. Novais, J.M. Pêgo, N. Sousa, J. Neves, Using mouse dynamics to assess stress during online exams, in: Proc. 10th Int. Conf. Hybrid Artif. Intell. Syst., HAIS, in: Lecture Notes in Computer Science, 9121, 2015, pp. 345–356, http://dx.doi.org/10.1007/978-3-319-19644-2_29.

[83] A. Kołakowska, Towards detecting programmers' stress on the basis of keystroke dynamics, in: Proc. 2016 Fed. Conf. Comput. Sci. Inf. Syst., FedCSIS, 2016, pp. 1621–1626, http://dx.doi.org/10.15439/2016f263.

[84] R. Bartels, T. Peçanha, HRV: a Pythonic package for heart rate variability analysis, J. Open Source Softw. 5 (51) (2020) 1867, http://dx.doi.org/10.21105/joss.01867.

[85] M. Malik, J.T. Bigger, A.J. Camm, R.E. Kleiger, A. Malliani, A.J. Moss, P.J. Schwartz, Heart rate variability: Standards of measurement, physiological interpretation, and clinical use, Eur. Heart J. 17 (3) (1996) 354–381, http://dx.doi.org/10.1093/oxfordjournals.eurheartj.a014868.

[86] M.P. Tarvainen, J.-P. Niskanen, J.A. Lipponen, P.O. Ranta-aho, P.A. Karjalainen, Kubios HRV – Heart rate variability analysis software, Comput. Methods Programs Biomed. 113 (1) (2014) 210–220, http://dx.doi.org/10.1016/j.cmpb.2013.07.024.

[87] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Adv. Neural Inf. Process. Syst. 30, NIPS, 2017, pp. 3146–3154.

[88] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K.J. Oedegaard, J. Tørresen, Mental health monitoring with multimodal sensing and machine learning: A survey, Pervasive Mob. Comput. 51 (2018) 1–26, http://dx.doi.org/10.1016/j.pmcj.2018.09.003.

[89] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Adv. Neural Inf. Process. Syst. 30, NIPS, 2017, pp. 4768–4777.

[90] A. Alberdi, A. Aztiria, A. Basarab, D.J. Cook, Using smart offices to predict occupational stress, Int. J. Ind. Ergon. 67 (2018) 13–26, http://dx.doi.org/10.1016/j.ergon.2018.04.005.

[91] W. Sanchez, A. Martinez, Y. Hernandez, H. Estrada, M. Gonzalez-Mendoza, A predictive model for stress recognition in desk jobs, J. Ambient Intell. Hum. Comput. (2018) http://dx.doi.org/10.1007/s12652-018-1149-9.

[92] N. Banholzer, S. Feuerriegel, E. Fleisch, G.F. Bauer, T. Kowatsch, Computer mouse movements as an indicator of work stress: Longitudinal observational field study, J. Med. Internet Res. 23 (4) (2021) e27121, http://dx.doi.org/10.2196/27121.

[93] R. Bavaresco, J. Barbosa, H. Vianna, P. Büttenbender, L. Dias, Design and evaluation of a context-aware model based on psychophysiology, Comput. Methods Programs Biomed. 189 (2020) 105299, http://dx.doi.org/10.1016/j.cmpb.2019.105299.