Stress detection in computer users from keyboard and mouse dynamics

Stress detection in Computer Users from Keyboard and Mouse Dynamics

Lucia Pepa, Antonio Sabatelli, Lucio Ciabattoni, *Member, IEEE,* Andrea Monteriù, *Member, IEEE*
Fabrizio Lamberti, *Senior Member, IEEE,* and Lia Morra, *Senior Member, IEEE*

***Abstract*—Detecting stress in computer users, while technically challenging, is of the utmost importance in the workplace, especially now that remote working scenarios are becoming ubiquitous. In this context, cost-effective, subject-independent systems are needed that can be embedded in consumer devices and classify users' stress in a reliable and unobtrusive fashion. Leveraging keyboard and mouse dynamics is particularly appealing in this context as it exploits readily available sensors. However, available studies are mostly performed in laboratory conditions, and there is a lack of on-field investigations in closer-to-real-world settings. In this study, keyboard and mouse data from 62 volunteers were experimentally collected in-the-wild using a purpose-built Web application, designed to induce stress by asking each subject to perform 8 computer tasks under different stressful conditions. The application of Multiple Instance Learning (MIL) to Random Forest (RF) classification allowed the devised system to successfully distinguish 3 stress-level classes from keyboard (76% accuracy) and mouse (63% accuracy) data. Classifiers were further evaluated via confusion matrix, precision, recall, and F1-score.**

***Index Terms*—Stress classification, machine learning, keyboard, mouse, in-the-wild study.**

I. Introduction

PROVIDING computer-based systems with the capability to recognize emotions is an ongoing subject of study. Should consumer devices like, e.g., laptops, smartphones, in-car entertainment systems and home appliances be capable to achieve an accurate reading of individuals' affective states, they could make appropriate decisions about how to interact with them, and adapt system's responses accordingly [1]. Applications are plentiful in fields like human-computer interaction, robotics, entertainment, learning, and healthcare.

Among emotional states that could be tackled, there is one that deserves a special attention, given the key role that it plays in work environments and for human health [22]: stress. Stress is a physiological response to a situation perceived to be challenging or threatening. While moderate levels of stress can be actually beneficial to work performance, chronic stress has been shown to be highly detrimental. Chronic stressors may lead to burnout, a growing concern in both western and developing countries with an estimated lifetime prevalence of 4% [40]. Evidence shows that workers make more errors when overly stressed, leading to a loss of productivity and, in the case of critical infrastructures, potentially fatal consequences [2]. Furthermore, stress may negatively impact the immune and cardiovascular systems [38]. Adding to this situation, the COronaVIrus Disease 2019 (COVID-19) outbreak led to a massive shift towards a Working From Home (WFH) operating modality, and public announcements by major tech companies are sparking a debate on the potential opportunities and perils of resorting to WFH on a permanent basis. In fact, despite its appeal, WFH may expose workers to new forms of stress and burnout, as the lines between professional and personal lives become blurry and workers struggle to preserve healthy boundaries between the two [39].

It is therefore crucial to equip both workers and managers alike with tools to enable proper stress management in a remote workforce, starting with methods to detect stress and other emotional states based on users' observation [28]. In the last years, different approaches have been investigated for stress detection [37], [28]. Even though some of them achieved quite impressive results, there are still serious problems limiting their applicability. First, most of the proposed methods rely on sensors directly attached to the users' skin or body [22], or use external recording sensors such as webcams, microphones, or even thermal cameras [14].

Both methods have side effects: first, users are aware of being monitored, which could alter their affective states and be itself a source of additional stress; second, it is unlikely that users can wear or use monitoring devices continuously during everyday activities [28]. Specialized hardware can be expensive and is unlikely to become commonplace in the short term or in a WFH scenario. Last but not least, both raise major privacy concerns, especially in a work environment.

Thus, the challenge appears to be the development of cost-effective, subject independent systems that can be embedded in consumer devices and that are able to detect users' stress in a reliable and unobtrusive fashion. In this paper, a possible solution to this challenge is proposed by leveraging the analysis of *keystroke and mouse dynamics* (K&MD). Many workers use a computer on a daily basis; thus, this solution would not require dedicated hardware, could be readily deployed in a traditional office or WFH setting, and would have minimal risks from a privacy viewpoint. Furthermore, affective states evaluation could be readily integrated in the work environment, e.g., to remind users to take pauses when overworked or overstressed.

Even though K&MD-based methods were proven suitable to identify several emotions with good performance [7], [21], [36], their practical application is still an open problem.

L. Ciabattoni, A. Monteriù, L. Pepa are with Università Politecnica delle Marche, Ancona, 60131, Italy (email: l.ciabattoni@staff.univpm.it; a.monteriu@staff.univpm.it; l.pepa@univpm.it)

A. Sabatelli is with Revolt SRL, Ancona, 60131, Italy (email: antonio.sabatelli@revoltsrl.it)

F. Lamberti and L. Morra are with Politecnico di Torino, Torino, 10129, Italy (email: fabrizio.lamberti@polito.it; lia.morra@polito.it)

Previous studies were mostly conducted in laboratory conditions, and there is a lack of on-field studies closer to actual professional settings [28]. In-the-wild studies face difficulties in inducing the intended stress levels, as well as collecting and labelling data, as the experimenter cannot directly interact with the subjects.

The present work tries to address the above challenges by designing a stress classification method based on K&MD that leverages real-world, in-the-wild data acquired in an uncontrolled setting resembling traditional office or WFH scenarios.

Specifically, the contribution of this paper is threefold:

- a web-based stress induction setup for collecting K&MD data *in the wild*: users are asked to engage in several tasks, representative of various computer-based activities, under different stressful conditions;
- fine-grained 3-level stress detection based on a variety of K&MD features;
- a cross-subjects validation design: while most previous works evaluated their algorithms through subject-dependent validations, which ensures higher accuracy [28], cross-subject validation is essential to quantify algorithm robustness, especially prior to its deployment in production environments [37].

This work extends a preliminary investigation [36] that demonstrated the feasibility of these objectives through a controlled study of stress detection, where just 2 classes were considered.

II. Background

A. Overview of stress detection techniques

Stress manifests itself in a plurality of ways which can be broadly classified as *psychological*, *behavioural* and *physiological*. While psychological effects may be evaluated through direct interaction with the user, e.g., through questionnaires or chatbots, stress detection is most commonly performed by detecting behavioral and physiological alterations through a variety of sensors, briefly categorized in Fig. 1.

Stress-related *physiological* processes, mediated by the autonomic nervous system, are largely involuntary changes in cardiovascular, muscular and electrodermal activity, respiratory rate, skin temperature, and eye movements [22]. They can be observed using a variety of sensors including *wearable sensors* [22], [4], [9] and, less commonly, eye tracking devices [12], [15] and thermal infrared imaging [22]. Recent advances in wearable sensors allows to record physiological signals in an increasingly unobtrusive, yet accurate fashion, yielding reliable and accurate stress measurements. Nonetheless, as users may not be willing to wear or use monitoring devices constantly, it is important to investigate complementary strategies.

Behavioral alterations in response to stress include bodily gestures (e.g., facial expressions, body pose) [13], [10] and speech [14], which can be detected using cameras, microphones and 3D cameras (e.g., Kinect) in combination with computer vision and speech analysis algorithms [14]. Another important line of research is detecting stress from daily life activity, such as eating [41], computer interactions [36], [5], [35], or driving. Besides the cost of deploying such sensors, there are significant privacy and acceptability issues associated with constantly recording subjects. In contrast, analyzing naturally occurring interactions with electronic devices, such as smartphones [17] and computers [7], [1], [36], [20], does not require additional and potentially intrusive hardware.

Another important distinction is related to the *setting* in which stress detection is carried out, e.g. during everyday home activities [4], [9], working [8], [17], [5], driving, and in outdoor places [4], [9]. Although some detection methods can target more than one setting, special-purpose approaches are much more common. We focus here in particular on professional and office environments. In [5], various technologies for monitoring office workers' emotions, including stress and mental load, were compared, and mouse and keyboard obtained the highest scores in most of the analyzed dimensions, including use of common hardware, cost-effectiveness, intuitiveness, availability and privacy compliance. Nonetheless, in another review on automatic stress recognition for office environments [35], only a handful out of the two hundred references cited in that work was actually based on K&MD. Additionally, lower accuracy is generally obtained compared to methods based on computer vision and wearable devices.

B. Stress detection from K&MD

Early studies of K&MD focused on simple *tasks* like password entering. When dealing with more sophisticated tasks (in terms of interactions), many factors may influence typing rhythm or mouse movements, including individuals' age, gender, handedness, skills, physical and mental state, and familiarity with the task, as well as external conditions, like hardware or software used, presence of disturbing elements, etc. [26]. Unsurprisingly, this fact pushed researchers to work under *laboratory controlled* conditions or, alternatively, to simplify on-field studies by focusing on specific tasks, such as computer programming [23], [6], [7]. An exception is reported in [1], in which participants of a user study were simply requested to carry out their usual daily activities (like, e.g., using a word processor, or an email application) and to rate their emotions from time to time. The main drawback of this approach is the lack of control on the actual emotions tested or the number of collected samples.

Both on-field and laboratory studies have to deal with the fundamental issue of *how to induce and rate affective states* [33]. A common solution is to rely on ad hoc tasks, often drawn from psychological literature, to raise participants' memory load, irritation, anxiety, pressure, or similar cognitive stress states. Participants are requested, e.g., to perform mental calculations [16], [26], play math- or logic-related games (e.g., Tower of Hanoi) [18], or stressing exercises (e.g., Stroop's color-word interference test [11]) [15], [13]), remember a number of words or digits (n-back memory task) [34], answer questions about a given clip or text [21], etc. A smaller number of works leveraged tasks closer to everyday activities, like transcribing a text [19] or searching an item in a website [29].

Different stress levels can be inducted by varying task complexity [12] and/or introducing external stressors [9]. For instance, the characteristics of the test environment can be

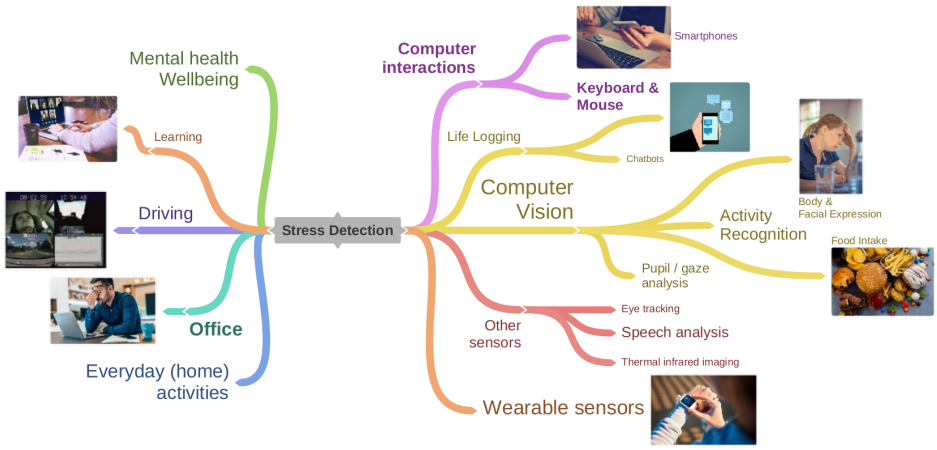

Fig. 1: Taxonomy of stress detection approaches.

changed from comfortable, to neutral and stressful through relaxing music, silence, or loud noise [24]. Other stressors include varying level of guidance [6], introducing time constraints, random disturbing events (faults, interruptions, etc.), monetary compensation [19] and social pressure [20], [17]. These stressors are similar to those typical of work environments, such as dealing with constant noise or interruptions or meeting strict deadlines.

Data labeling, i.e., rating of perceived (level of) emotions is another critical factor, in particular for on-field studies. While in some cases, external raters or recordings are used [31], the most frequent option is self-rating [1], [30]. The latter approach is also more suited to in-the-wild data collection, where additional sensors are not easy to deploy.

A dimension that further distinguishes works in this field concerns the diversity of *features* used. This is partly due to the fact that keyboard- and mouse-based features may be task-specific [26]. Early research focused on keyboard activity, as password typing had been extensively studied for authentication purposes [3]. More recent work [16] leveraged keystroke pause length, time per keystroke, time between keystrokes, and frequency of deletion and navigation keys [16], as well as mouse speed and directions [32].

Results showed that, in laboratory conditions, it is feasible to classify stress conditions with comparable accuracy to other affective states (75% with k-NN). Better results have been only obtained by using ad hoc hardware or combining K&MD with other sensing techniques [27].

By moving from the above review, several research gaps were identified and addressed in this study. First, few works addressed in-the-wild setups [37], [22], and even fewer carried out a proper stress classification. Some had to fall back to a more general valence [23] or emotion [1], [6] classification, mainly because of the lack of data. Khan et al. [7] performed just a regression analysis on data collected in-the-wild, suggesting the potential for a future stress detection. Many works focused on 2-class stress detection; however, a stress classification of at least 3 levels would be closer to clinical evaluations and more useful in real-life applications [38]. Finally, algorithms validation can be performed through within-subject or between-subjects cross validation methods. Within-subject validation often leads to better results, but the ability of an algorithm to generalize over unseen users is crucial to quantify its robustness and suitability to real scenarios [37], [14], [9]. Additionally, training user-specific classifiers would require a great amount of data, thus limiting applicability in many real-world applications [23], [6]. The need to investigate between-subjects multiclass stress classification is much more evident for in-the-wild studies, since previous literature did not address this issue.

III. Experimental Protocol for Stress Induction

A. Subjects Recruitment

Since our study focuses on stress detection in-the-wild, few constraints were set in recruiting subjects. In particular, an invitation was sent to students and teaching staff at the authors' universities, who were asked to extend the invitation to relatives and friends. The study was conducted during the COVID-19 outbreak, when subjects were mostly working remotely. A total of 62 subjects were recruited. All of them had to be at least 18 years old (28 on average, $\sigma = 8$, 40 males and 22 females), native Italian speakers, and use computers daily. Subjects with previous history of cardiac, neurological or anxiety disorders, color blindness, or prescription drugs for sleep disorders were asked to self-exclude from the study. Subjects had to confirm that they had not consumed any alcohol or caffeine the day of the experiment, and any psychoactive drug within the 48 hours preceding the experiment.

B. Equipment

The experimental protocol was administered through a purpose-built Web application, allowing participants to complete the test whenever they wanted, from their homes or offices, and using their own equipment (keyboard, mouse, screen, etc.). The application, which could be accessed through a link provided with the invitation (https://www.revoltsrl.it/stress/#/welcome), was implemented using Angular, a TypeScript based open-source front-end framework for web development. The application is responsible both for administering

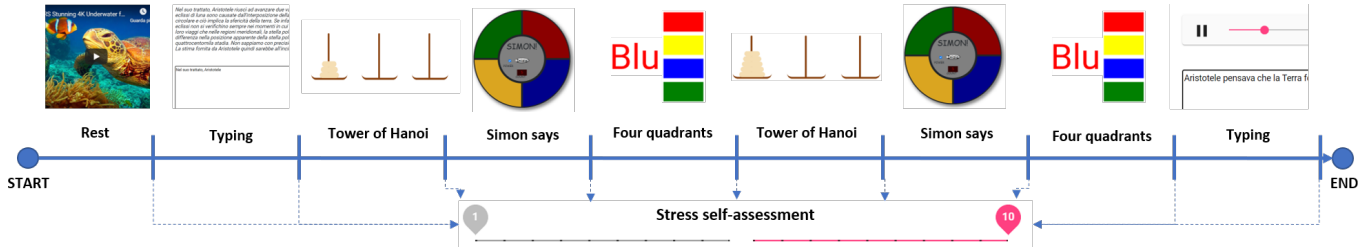

Fig. 2: Scheme of the experimental protocol. After an initial rest phase, the procedure comprises a sequence of 8 tasks (4 tasks, each repeated twice with increasing difficulty). After each task, the subject assessed his/her stress level on a 1-to-10 scale.

the tasks, as well as for collecting data concerning keyboard and mouse operations, as detailed in Section IV-A.

C. Procedure and Tasks

The experimental protocol was designed to induce stress by performing several computer tasks while distracted by sounds and other disturbances. Prior to the experiment, subjects were informed about the number of tasks, that each task had been designed to mimic common work or leisure activities, and that after each one they had to self-evaluate their perceived stress level on a scale from 1 (low) to 10 (high).

The experimental protocol started with a *rest phase* in order to set a baseline for stress measurement. Subjects were asked to watch a relaxing movie on underwater nature, included in the application, for five to ten minutes, or until they felt as relaxed as possible. After rating their perceived stress level, they were allowed to start the procedure.

The procedure comprises four tasks, each executed twice with increasing difficulty (Fig. 2):

- Text typing (easy), copying a short text;
- Tower of Hanoi (easy), with 3 disks;
- Simon Says game (easy), sequence of 5 sounds;
- Four-quadrant test (easy);
- Tower of Hanoi (difficult), with 5 disks;
- Simon Says game (difficult), sequence of 10 sounds;
- Four-quadrant test (difficult);
- Text typing (difficult), transcribing a dictated text.

The tasks were selected for their potential to raise cognitive load and anxiety based on a preliminary study conducted by the authors [36], as well as previous studies [22], [35].

In the first text typing task, which resembles everyday office activities [19], provided instructions stated that it was neither an accuracy contest nor a race; subjects had to type at a normal pace, and take the time needed for fixing possible mistakes. Average duration of this task was 339 seconds ($\sigma = 129$).

In the second task, subjects were requested to solve a simple Tower of Hanoi game (like, e.g., in [18]). This is a well-known problem-solving task in experimental psychology, being a relatively straightforward puzzle that requires very simple instructions and no additional domain knowledge. It consists of 3 rods and a number of disks of different sizes, which can only be moved on top of smaller disks. As shown in Fig. 2, the game starts with the disks properly stacked in the left rod. The objective is to move the entire stack to the very right rod by obeying the above constraints. With 3 disks, the puzzle can be solved in 7 moves. No time constraints were set. On average, subjects needed 71 seconds and 15 moves to solve the puzzle ($\sigma = 38$ and $\sigma = 5$, respectively).

The third task was aimed at collecting mouse data by inducing a given stress level through a n-back task. In order to approximate a common computer task, we implemented a web-based version of Simon Says, a well-known electronic memory game [34]. The game requires the subjects to repeat a sequence of sounds, increasing the length by one every time they succeed, until the maximum number of sounds is reached. In case of errors, the sequence is repeated. On average, subjects made 4 errors ($\sigma = 3$) and spent 77 seconds ($\sigma = 48$) on this task.

The fourth task modeled the impact that interferences can have on task execution, increasing subjects' reaction time and perceived stress. In this four-quadrant task [2] (a variant of the Stroop test [11]), subjects are shown the names of several colors displayed in differently colored fonts (e.g., the word "red" in a yellow font). Subjects have to click on the colored quadrant which corresponds to the spelled word ("red" in our example), instead of the font color (yellow in our example). In previous studies [2], subjects had instead to click on the quadrant corresponding to the font color and ignore the word. Preliminary experiments showed that higher level of stress could be induced by the chosen implementation [36]. Subjects had a maximum of 3 seconds for clicking on the right quadrant: after that, a new word and color were generated. Every time they did not click the right quadrant in time, a strong and unpleasant buzz sound was played. The task lasted 90 seconds on average ($\sigma = 54$).

In the fifth task, the Tower of Hanoi game was used again. However, in order to induce much higher levels of stress, 5 disks were used (requesting at least 31 moves). Moreover, two different stressors were added, namely a disturbing tick-tock sound and a timer (set to 300 seconds). On average, subjects needed 215 seconds ($\sigma = 82$) and 80 moves ($\sigma = 34$).

Similarly, in the sixth task, the Simon Says game was used again, with sequences of 10 sounds. On average, subjects made 10 errors ($\sigma = 9$) and spent 146 seconds ($\sigma = 52$) on this task. In the seventh step, the four-quadrant task was repeated with the addition of a tick-tock sound. Time limit for clicking the correct quadrant was also reduced to 2 seconds. Subjects employed on average 135 seconds ($\sigma = 43$) to complete it.

In the last task, subjects were requested to type a dictated text. Dictation speed was rather high, and it was not possible to pause it. Subjects were requested to type at their fastest

pace, trying to fix typos and errors if possible. The duration of this task was 320 seconds. At the end of the experiment, collected data were saved on the computer, and subjects were requested to send them for dataset creation and processing.

IV. Proposed Method for Stress Classification

A. Data Collection

The data collected during the experimental protocol are divided in 3 categories: stress self-assessment, keyboard data, and mouse data. Self assessment data consists of self-reported measures of the perceived stress level (on a 1 to 10 scale) collected after each task of the experimental protocol. Keyboard data were acquired during the 2 text typing tasks. For each keystroke, the software records the character typed, if it was a keyup or keydown event (boolean value), the duration of the pressure (ms) and a timestamp (ms). Mouse data were acquired during the Tower of Hanoi, Simon Says, and four-quadrant tasks and included, for each click, the mouse coordinates (x and y), the presence of a press or release (boolean value, for both the left and right button), the click duration and dwell time (in ms), and a timestamp (in ms). All the data were registered by the Angular application, exported as CSV files and imported in MATLAB for data analysis and classification.

B. Feature Extraction

A sliding window of 5 seconds without overlap was applied on keyboard and mouse data, and feature extraction was performed on each window. For features that yield an array of values for each time window, the maximum, minimum, mean, standard deviation (std), and point-to-point (ptp) variation (difference between maximum and minimum) were extracted, for a total of 5 features. Other features were directly computed as a single value in the window. In the following, feature categories are referred to as either "array" or "single value".

In summary, 15 keyboard features were computed [16], [25]:

- key dwell time (array): press-to-release time of each key (ms);
- key down-to- down time (array): time elapsed from the press of one key to the press of the next key (ms);
- key velocity (single value, mean): number of keys pressed per second;
- latency time (single value, mean): time elapsed from a key release to the press of the next key (ms);
- number of backspaces (single value, count);
- number of key pressed (single value, count);
- key press (single value, percentage): amount of the window with at least one key pressed.

From mouse data, 22 features were computed [32]:

- mouse velocity (array): change in position per second;
- mouse acceleration (array): variation of velocity per second;
- mouse inactivity (single value, count): time for which the user is not moving the mouse (ms);
- number of clicks (single value, count);
- click dwell time (array): duration of each click (ms);
- click distance (array): time distance between two consecutive clicks (ms).

All the features were validated in previous literature, as well as in a preliminary experiment in laboratory conditions [36]. Some features may assume an invalid value in given windows (e.g., click distance needs at least two clicks in order to be significant): these windows were excluded from the analysis.

C. Stress Classification Algorithm

Keyboard and mouse are rarely used at the same time when working at a computer; hence, different classifiers were built in order to predict the stress level depending on what device the participant was using. Each time windows was labelled as *low* (1–3), *medium* (4–7) or *high* (8–10) stress. After *min-max* normalization, feature selection was performed based on Neighborhood Component Analysis (NCA). In previous work, best performance were achieved when limiting the analysis to the most discriminative features [36]. Building on (and further extending) our preliminary investigation [36], several ML techniques were compared in order to build the classifiers: k-Nearest Neighbour, Support Vector Machines, Decision Trees, and Random Forest (RF). Since RF reached the best results, the remainder of this work will focus on this method.

In order to deal with sparse and inaccurate labeling deriving from on-field data, a Multiple Instance Learning (MIL) approach was applied. MIL is a semi-supervised learning technique where the task is learned given labelled groups, or "bags", each containing multiple training samples. Since MIL does not assume complete knowledge of training labels, it is particularly suited to in-the-wild data analysis. In our case, all the time windows from the same task share the same label; the classifier will learn through approximately classified time intervals (bags) rather than individual instances (single time windows). In this paper, Majority-Voting RF is selected as MIL extension of a RF classifier. Bags length was set to 90 seconds with an overlap of 50%.

A subject-independent 5-fold cross validation was adopted to test both classifiers: 80% of the participants were used for training, the remaining 20% for testing. The classifier is thus tested on never-seen participants, which is a condition close to real applications, as discussed in Section II. Classifiers performance was evaluated on validation bags via confusion matrix, accuracy, precision, recall, and F1-score. Performance scores are averaged over the 5 folds.

V. Results

All the participants successfully completed all the planned tasks, for a total of 496 tasks. Task recordings that were empty (38 tasks, 7.7%), clearly shorter than the minimum time required to complete the task (8), or much longer than the maximum plausible time (8) were excluded. After the exclusion of invalid trials, a total of 411 (100 low, 219 medium, 92 high) and 429 (120 low, 222 medium, 87 high) bags were extracted from mouse and keyboard data respectively.

Participants' stress levels were compared between the easy and difficult version of each task (Fig. 3) and between each task and the rest phase. At one-way non-parametric ANOVA, median differences were statistically significant ($p < 0.05$) for all the tasks except the easy Tower of Hanoi and the rest phase.

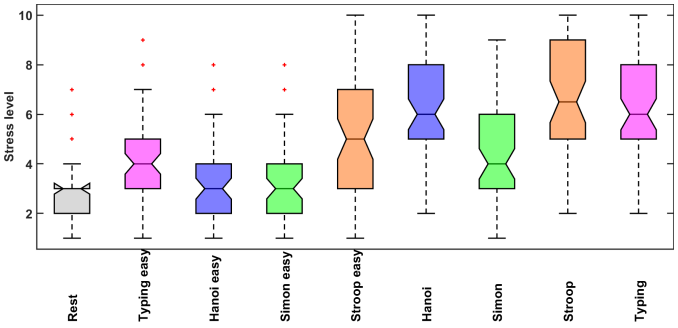

Fig. 3: Stress levels self-assessed by participants for the tasks in the experimental protocol. The easy and difficult versions of the same task are displayed in the same color (typing tasks: purple; Tower of Hanoi tasks: blue; Simon Says tasks: green; four-quadrant tasks: orange), rest phase is in gray.

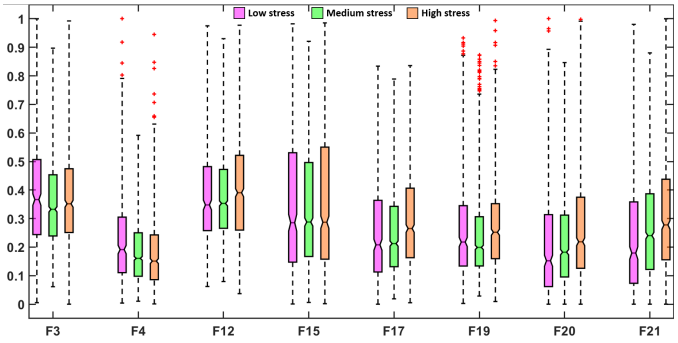

Fig. 4: Distribution of the most discriminative mouse features according to NCA feature selection. F3, F4: mouse velocity (mean, std). F12, F15: click dwell time (maximum, std). F17, F19, F20, F21: click time distance (maximum, mean, std, ptp).

The mouse most discriminative features (Fig. 4), as selected by the NCA, were: mouse velocity (mean, std), click dwell time (maximum, std), and click distance (maximum, mean, std, ptp). The keyboard most discriminative features (Fig. 5) were: key dwell time (maximum, minimum, std), key velocity, key down-to-down time (minimum, std), number of key pressed, key press percentage, and latency time.

Classification accuracy reached 63% and 76% for mouse and keyboard classifiers, respectively. Fig. 6 shows confusion matrices for the keyboard (Table 6a) and mouse (Table 6b) classifiers. Columns are the true classes, rows the predicted ones. The considered classes, i.e., low, medium, and high stress are indicated with letters L, M, and H respectively. Classification performance in terms of recall, precision and F1-score is presented in Table I.

	Recall			Precision			F1-score		
	L	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**
K	0.57	0.87	0.47	0.4	0.85	0.64	0.47	0.86	0.54
M	0.24	0.77	0.75	0.42	0.73	0.75	0.3	0.69	0.75

TABLE I: Recall, precision, and F1-score of the keyboard (K) and mouse (M) classifiers for the 3 classes, low (**L**), medium (**M**), and high (**H**) stress.

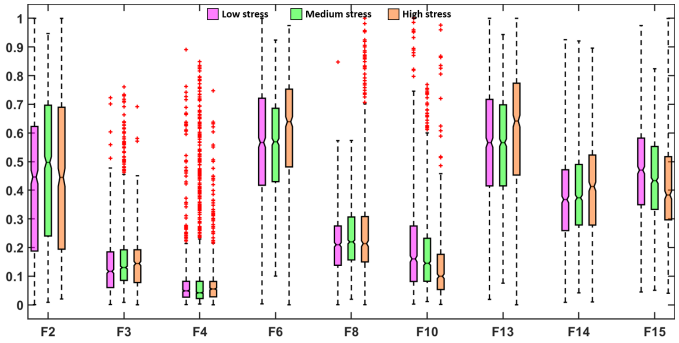

Fig. 5: Distribution of the most discriminative typing features according to NCA feature selection. F2, F3, F4: key dwell time (minimum, mean, std). F6: key velocity. F8, F10: key down-to-down time (minimum, std). F13: number of keys. F14: key press percentage. F15: mean latency time.

Pred. \ True	**L**	**M**	**H**
L	4	3	3
M	3	47	5
H	0	4	7

(a)

Pred. \ True	**L**	**M**	**H**
L	5	5	2
M	15	30	3
H	1	4	15

(b)

Fig. 6: Confusion matrix for (a) keyboard and (b) mouse classifier. Columns: true classes, rows: predicted classes.

VI. Discussion

Statistical analysis of participants' stress self-assessment revealed that the selected tasks were able to increase the stress level from the rest condition, as well as from the easy to the difficult versions of each task. At the same time, it showed that the stress level varied from one task to the other, as confirmed by post-hoc interviews with participants. Indeed, the experimental protocol was devised to re-create a scenario similar to a common working day, in which different tasks need to be accomplished under varying stress levels. Eventually, the experimental protocol induces an overall increase in perceived stress level which can be correctly detected by machine learning techniques.

Differences in tasks, data, classes, and algorithms make it difficult to directly compare obtained results with existing literature. However, it is worth summarizing the current and previous contributions concerning affective states or stress detection using K&MD, as proposed in Table II, which reports the number of participants, setting, sensor, algorithm, number of classes and accuracy. The current work is the only performing a ternary stress classification using in-the-wild data collected from a wide sample (62 participants) and reaching an accuracy above 75% (using KD). Several discriminative features were found significant also in previous works, such as key dwell time [1], [16], [21], key down-to-down [1], [21], [23], key latency [1], [21], number of keys or keys rates [1], [16], and mouse velocity [20], [23]. Decision trees and RF were also investigated in some previous works leading to better performance with respect to other algorithms [21], [1]. However, the introduction of a MIL

Paper	Subj.	Setting	Sensor	Algorithm	Classes	Accuracy
[1]	12	real	KD	C4.5	2	<75%
[16]	24	lab.	KD	KNN	2	75%
[20]	18	lab.	MD	threshold	2	78%
[21]	35	lab.	K&MD	RF	2	69%
[23]	14	lab.	K&MD	FFNN	3	52.9%
[28]	25	lab.	K&MD	SVM	2	65.5%
this	62	real	K&MD	RF+MIL	3	76.3%

TABLE II: Comparison with the literature. Acronyms: FFNN (Feed-Forward Neural Network), KNN (K-Nearest Neighbour), SVM (Support Vector Machine), RF (Random Forest), (MIL) (Multiple Instance Learning).

approach possibly contributed to better results by mitigating the effect of inaccurate labeling, typical of natural settings.

While in line with previous literature, our results also showed some weaknesses, especially in the prediction of the low stress class (F1-score is equal to 0.47 for the keyboard, and 0.3 for the mouse). In contrast, for both the medium and high classes, precision and recall are comparable, as indicated by the F1-score (keyboard, 0.84 for medium stress and 0.54 for high stress, mouse, 0.69 for medium stress and 0.75 for high stress) and well above the chance level. However, an error in classifying the low stress class has a much lower impact that an error in classifying a high stress class. These problems were also mentioned in previuos works: the difficulty to obtain high performance from K&MD is known in on-field studies [1], and it was considered a drawback of the lack of control on induced affective states and data loss.

Overall, based on the above considerations the outcomes of this study appear to be promising and particularly relevant for future developments, especially considering that results from real-world, in-the-wild setups are generally regarded as much more informative than those from controlled setups [37], [22].

Some limitations of the proposed approach, such as data loss, are typical of in-the-wild experiments. Other encountered limitations were documented also in controlled studies, like labeling uncertainty, class imbalance (due to the difficulty to solicit high stress levels), and task selection (for instance, the Simon Says and the four-quadrant tasks tend to generate specific mouse patterns that may not generalize for different tasks). This last aspect could be further investigated by including tasks that are less clinically relevant as stressors, but that are more similar to everyday computer activities.

VII. Conclusion

The aim of this work was to reach a subject-independent, multiclass stress classification in computer users in an uncontrolled environment through a non-intrusive, non-invasive, and cost-effective solution. To this purpose, data generated by common multimedia input peripherals were collected in-the-wild from 62 subjects using their own computer-based equipment. MIL applied to a RF algorithm reached the best results in classifying 3 stress levels. While confirming some of the limitations known in the literature, the findings of this study contribute at shedding further light on a challenging, though extremely important goal for this field of research.

Future works will explore how K&MD could be combined with other stress detection methods (e.g., vision-based methods or wearable sensors). The proposed methodology can be extended by exploring other tasks, more closely related to real-life tasks, and by exploring inter-subject as well as cross-subject designs. An open challenge is how to disentangle variations in K&MD patterns related to the task to those due to the users' stress response. A further research direction to explore is how stress detection can be leveraged to enhance human-computer interface, e.g. by adapting the system behavior according to the user's emotional states, or providing feedback to the users in order to increase their awareness of their cognitive and mental state. A computer or mobile device may integrate data acquired by multiple IoT devices and wearable sensors at different times of the day to build an accurate, fine-grained and dynamic picture of the user's cognitive and emotional state.

References

[1] C. Epp, M. Lippold, and R. Mandryk, "Identifying emotional states using keystroke dynamics," In: Proc. Conference on Human Factors in Computing Systems, pp. 715–724, 2011.

[2] S. H. Lau, "Stress detection for keystroke dynamics," MA thesis. Carnegie Mellon University, School of Computer Science, 2018.

[3] D. L. Salil, and P. Banerjee, "Biometric authentication and identification using keystroke dynamics: A survey," *Journal of Pattern Recognition Research*, vol. 7, pp. 116–139, 2012.

[4] Y. S. Can , B. Arnrich , and C. Ersoy, Stress detection in daily life scenarios using smart phones and wearable sensors: A survey, *Journal of Biomedical Informatics*, vol. 92, 2019.

[5] D. Carneiro, P. Novais, J.C. Augusto, and N. Payne, New methods for stress assessment and monitoring at the workplace, *IEEE Transactions on Affective Computing*, vol. 10, no. 2, 2019.

[6] A. Kolakowska, Towards detecting programmers' stress on the basis of keystroke dynamics, In: Proc. of the Federated Conference on Computer Science and Information Systems, pp. 1621–1626, 2016

[7] I. A. Khan, W. P. Brinkman, and R. Hierons, Towards estimating computer users' mood from interaction behaviour with keyboard and mouse, *Frontiers of Computer Science*, vol. 7, pp. 943–954, 2013.

[8] A. Belk, D. Portugal, P. Germanakos, J. Quintas, E. Christodoulou, and G. Samaras, A computer mouse for stress identification of older adults at work, In: Proc. 1st Int. Workshop on Human Aspects in Adaptive and Personalized Interactive Environments, pp. 1–4, 2016.

[9] L. Rachakonda, S. P. Mohanty, E. Kougianos, and P. Sundaravadivel, Stress-Lysis: A DNN-integrated edge device for stress level detection in the IoMT, *IEEE Transactions on Consumer Electronics*, vol. 65, no. 4, 2019.

[10] I. Lefter, G. J. Burghouts, L.J.M. Rothkrantz, Recognizing stress using semantics and modulation of speech and gestures, *IEEE Transactions on Affective Computing*, vol. 7, no. 2, 2016.

[11] J. R. Stroop, Interference in serial verbal reactions, *Journal of Experimental Psychology*, vol. 18, pp. 643–661, 1935.

[12] C. Jyotsna, and J. Amudha, Eye gaze as an indicator for stress level analysis in students, In: Proc. Int. Conf. on Advances in Computing, Communications and Informatics, pp. 1–6, 2018.

[13] G. Giannakakis, D. Manousos, V. Chaniotakis, M. Tsiknakis, Evaluation of head pose features for stress detection and classification, In: Proc. IEEE EMBS International Conference on Biomedical & Health Informatics, pp. 406–409, 2018

[14] A. Riera, A. Soria-Frisch, A. Albajes-Eizagirre, P. Cipresso, C. Grau, S. Dunne, and G. Ruffini, Electro-physiological data fusion for stress detection, *Studies Health Technol. Informat.*, vol. 181, pp. 228–232, 2012.

[15] F. Mokhayeri, M.-R. Akbarzadeh-T, and S. Toosizadeh, Mental stress detection using physiological signals based on soft computing techniques, In: Proc. 18th Iranian Conference on Biomedical Engineering, pp. 232–237, 2011.

[16] L. M. Vizer, L. Zhou, and A. Sears, Automated stress detection using keystroke and linguistic features: An exploratory study, *International Journal of Human-Computer Studies*, vol. 67, no. 10, pp. 870–886, 2009.

[17] M. Ciman, and K. Wac, Individuals' stress assessment using human-smartphone interaction analysis, *IEEE Transactions on Affective computing*, vol. 9, no. 1, pp. 51–65, 2018.

[18] L. Ciabattoni, F. Ferracuti, S. Longhi, L. Pepa, L. Romeo, and F. Verdini, Real-time mental stress detection based on smartwatch, *IEEE International Conference on Consumer Electronics*, pp. 1–2, 2017.
[19] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski, Under pressure: Sensing stress of computer users, In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 51-60, 2014.
[20] T. Kowatsch, F. Wahle, and A. Filler, Design and lab experiment of a stress detection service based on mouse movements, In: Proc. 11th Mediterranean Conference on Information Systems, pp. 1–17, 2017.
[21] R. Shikder, S. Rahaman, F. Afroze, and A. A. Islam, Keystroke/Mouse usage based emotion detection and user identification, In: Proc. International Conference on Networking, Systems and Security, pp. 1–9, 2017.
[22] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, Review on psychological stress detection using biosignals, *IEEE Transactions on Affective Computing*, in press
[23] H. Liu, O. Noel, N. Fernando, and J. C. Rajapakse, Predicting affective states of programming using keyboard data and mouse behaviors, In: Proc. 15th International Conference on Control, Automation, Robotics and Vision, pp. 1408–1413, 2018.
[24] D. E. Vargas Ligarreto, and D. López De Luise, Metrics design for keyboard and mouse: Assessing learning levels, In: Proc. Congress on Electronics, Electrical Engineering and Computing, pp. 1–4, 2017.
[25] K. Revett, F. Gorunescu, M. Gorunescu, M. Ene, S. Magalhaes, and H. Santos, A machine learning approach to keystroke dynamics based user authentication, *International Journal of Electronic Security and Digital Forensics*, vol. 1, no. 1, pp. 55–70, 2007.
[26] Y.M. Lim, A. Ayesh, and M. Stacey, Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic, In: Proc. Science and Information Conference, pp. 146–152, 2014.
[27] M. X. Huang, J. Li, G. Ngai, and H. V. Leong, StressClick: Sensing stress from gaze-click patterns, In: Proc. 24th ACM international conference on Multimedia, pp. 1395–1404, 2016.
[28] S. Koldijk, M. A. Neerincx, and W. Kraaij. Detecting work stress in offices by combining unobtrusive sensors, *IEEE Transactions on Affective Computing*, vol. 9, no. 2, 2018.
[29] A. van Drunen, E. L. van den Broek, A. J. Spink, and T. Heffelaar, Exploring workload and attention measurements with Log mouse data, *Behavior Research Methods*, vol. 41, no. 3, pp. 868–875, 2009.
[30] W. Maehr, eMotion: Estimation of the user's emotional state by mouse motions, VDM Verlag Dudweiler Landstr., Saarbrucken, Germany
[31] A. Althothali, Modeling user affect using interaction events, Master thesis, University of Waterloo, Canada, 2011.
[32] G. Tsoulouhas, D. Georgiou, and A.Karakos, Detection of learners' affective state based on mouse movements, *Journal of Computing*, vol. 3, no. 11, pp. 9–18, 2011.
[33] A. Kolakowska, A review of emotion recognition methods based on keystroke dynamics and mouse movements, In: Proc. 6th Int. Conference on Human System Interactions, pp. 548–555, 2013.
[34] N. Z. Gurel, H. Jung, S. Hersek, and O. T. Inan, Fusing near-infrared spectroscopy with wearable hemodynamic measurements improves classification of mental stress, *IEEE Sensors Journal*, vol. 19, no. 9, pp. 8522–8531, 2019.
[35] A. Alberdi, A. Aztiria, and A. Basarab, Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of Biomedical Informatics*, vol. 59, pp. 49–75, 2016
[36] L. Ciabattoni, G. Foresi, F. Lamberti, A. MonteriÙ and A. Sabatelli, "A Stress Detection System based on Multimedia Input Peripherals," 2020 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2020, pp. 1-2.
[37] S. S. Panicker, P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," Biocybernetics and Biomedical Engineering, Volume 39, Issue 2, 2019, Pages 444-469, ISSN 0208-5216.
[38] Sharma N, Gedeon T. Objective measures, sensors and computational techniques for stress recognition and classification: a survey. Comput Methods Programs Biomed. 2012;108(3):1287-1301.
[39] B.E. Ashforth, G.E. Kreiner, and M. Fugate, "All in a day's work: Boundaries and micro role transitions," Academy of Management review, vol. 25, no. 3, pp.472–491, 2020
[40] M.K. Wekenborg, B. von Dawans, L.K. Hill, J.F. Thayer, M. Penz and C. Kirschbaum, "Examining reactivity patterns in burnout and other indicators of chronic stress," Psychoneuroendocrinology, vol. 106, pp.195–205, 2019
[41] Laavanya Rachakonda, Saraju P. Mohanty and Elias Kougianos. "iLog: an intelligent device for automatic food intake monitoring and stress detection in the IoMT." IEEE Transactions on Consumer Electronics, vol. 66, no. 2, pp. 115–124, 2020

Lucia Pepa received in 2012 the Master degree in Electronic Engineering, and in 2016 the Ph.D. degree in E-learning – Technology Enhanced Learning from the Universita' Politecnica delle Marche (UNIVPM), Italy. She is currently postdoc researcher at UNIVPM, her primary research interests involve affective computing and movement analysis through consumer electronics devices.

Antonio Sabatelli received the M.Sc. degrees in biomedical engineering from Università Politecnica delle Marche, Italy, in 2019. Currently he is a software engineer at Revolt SRL, Ancona, Italy. His research interests include computational intelligence, biomedical signal processing, consumer electronics devices.

Lucio Ciabattoni received the M.Sc. and the Ph.D. degrees from Università Politecnica delle Marche, Italy, in 2010 and 2014. Currently he is an assistant professor at the Department of Information Engineering (DII), Università Politecnica delle Marche, and the chair of the IEEE Italy Section CE Society Chapter. His research interests include computational intelligence, AI, renewable energy solutions, consumer electronics devices.

Andrea Monteriù (S'04-M'06) received the M.Sc. degree in Electronic Engineering and the Ph.D. degree in Artificial Intelligence Systems from Università Politecnica delle Marche, Italy, in 2003 and 2006. He is now an associate professor at Università Politecnica delle Marche, and the Vice-Chair of the IEEE Italy Section CE Society Chapter. Monteriù's research interests mainly focus on the areas of fault diagnosis and fault tolerant control applied on robotic, unmanned and artificial intelligent systems.

Fabrizio Lamberti received his M.Ss. and his Ph.D. degrees in computer engineering from Politecnico di Torino, Italy, in 2000 and 2005. He is now a full professor at Politecnico di Torino. His research interests include computer graphics, human-machine interaction and intelligent systems. He is serving as Associate Editor for IEEE Transactions on Consumer Electronics, IEEE Transactions on Computers, IEEE Transactions on Learning Technologies, and IEEE Consumer Electronics Magazine.

Lia Morra received the M.Sc. and the Ph.D. degrees in computer engineering from Politecnico di Torino, Italy, in 2002 and 2006. Currently, she is senior post-doctoral fellow at the Dip. di Automatica e Informatica of Politecnico di Torino. She is serving as Associated Editor for the IEEE Consumer Electronics Magazine. Her research interests include computer vision, pattern recognition, and machine learning.