IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Cognitive Training and Stress Detection in MCI Frail Older People through Wearable Sensors and Machine Learning

F. DELMASTRO[1], F. DI MARTINO [2], AND C. DOLCIOTTI[3]

[1]IIT Institute, National Research Council of Italy (CNR), PI 56124 Italy (e-mail: franca.delmastro@iit.cnr.it)
[2]IIT Institute, National Research Council of Italy (CNR), PI 56124 Italy (e-mail: flavio.dimartino@iit.cnr.it)
[3]Dept. of Translational Research of New Technologies in Medicine and Surgery, University of Pisa, PI 56126 Italy (e-mail: c.dolciotti@gmail.com)

Corresponding author: F. Delmastro (e-mail: franca.delmastro@iit.cnr.it).

This research has been partially funded by the INTESA project, co-funded by the Tuscany Region (Italy) under the Regional Implementation Programme for Underutilized Areas Fund (PAR FAS 2007-2013) and the Research Facilitation Fund (FAR) of the Ministry of Education, University and Research (MIUR).

ABSTRACT Personalised training of motor and cognitive abilities is fundamental to help older people maintain a good quality of life, especially in case of frailty conditions. However, the training activity can increase the stress level, especially in persons affected by a chronic stress condition. Wearable technologies and m-health solutions can support the person, the medical specialist, and long-term care facilities to efficiently implement personalised therapy solutions by monitoring the stress level of each subject during the motor and cognitive training. In this paper we present a comprehensive work on this topic, starting from a pilot study involving a group of frail older adults suffering from Mild Cognitive Impairment (MCI) who actively participated in cognitive and motor rehabilitation sessions equipped with wearable physiological sensors and a mobile application for physiological monitoring. We analyse the collected data to investigate the stress response of frail older subjects during the therapy, and how the cognitive training is positively affected by physical exercise. Then, we evaluated a stress detection system based on several machine learning algorithms in order to highlight their performances on the real dataset we collected. However, stress detection algorithms generally provide only the identification of a stressful/non stressful event, which is not sufficient to personalise the therapy. Therefore, we propose a mobile system architecture for online stress monitoring able to infer the stress level during a session. The obtained result is then used as input for a Decision Support System (DSS) in order to support the medical user in the definition of a personalised therapy for frail older adults.

INDEX TERMS Cognitive training, m-Health, Physiological Data analysis, Stress Detection, Wearable Sensors, Decision Support System, Machine Learning

I. INTRODUCTION

TODAY, ageing represents an increasing phenomenon, involving even more complex health conditions due to the coexistence of multiple chronic diseases. This generates a decreasing trend in the quality of life of both older people and their caregivers, often flowing in *frailty* condition.
Nowadays, frailty is considered a geriatric syndrome, usually involving over 65 years old people, representing a state of vulnerability with increased risk of poor health outcomes, including falls, incident disability, hospitalization and mortality [1]. Common signs and symptoms of frailty are weight loss, fatigue, muscle weakness and reduced physical and mental performances [2]. This condition, in addition to a possible social isolation, can also contribute to increase the psycho-physical stress, which can negatively affect sleep quality, mood, and cognitive performances. [3]. The entire picture refers to a complex clustering of emotional symptoms and behavioural features that generates a chronic stress condition. In addition, several studies established a link between frailty and cognitive decline [4]. Specifically, cognitive decline in older people is currently defined as Mild Cognitive Impairment (MCI). MCI is frequently associated with physiological ageing, but it can often represent a transitional phase from cognitive changes or deficits of normal ageing to patho-

logical features found in neurodegenerative and vascular dementia [5]. These two conditions, which often coexist in older people, can represent a cumulative health risk factor [6]. Specifically, MCI can affect an important domain of cognitive processes called executive functions, which refer to the ability to execute and maintain goal-directed behaviour. Working memory, inhibitory control, and cognitive flexibility represent the core components of the executive functions [7]. Therefore, in order to slow down and reduce the risk of collapse of frailty, it is necessary to define personalised programs, both in terms of health monitoring solutions, and cognitive and motor rehabilitation, which are often suggested by gerontologists. However, personalisation should also take into account the level of stress that this type of programs can generate in each subject, trying to identify the ideal training condition in terms of both improvements and personal compliance. In fact, even though it has been demonstrated that a proper training helps maintain efficient cognitive and motor abilities, the relationship between the frailty status and the execution of specific training sessions has not been investigated yet [8], especially in terms of related stress.

Stress can be classified into two main categories: acute and chronic stress. Acute stress refers to a transient occurrence of a single stressing agent, whereas chronic stress refers to an ongoing difficulty or disease that may be a constant threat to the individual's life [9]. The human body copes with stress using two response systems: a quick response ('fight-or-flight') to acute stress following the sympathetic adrenal medullary (SAM) axis, and a slow response to chronic stress following the hypothalamic pituitary adrenal (HPA) axis [10]. When the human brain faces a stressor, i.e., an internal or external agent that alters the body's homeostasis, the activation of the system response aims to restore the body's internal balance through physiological and behavioural adaptive responses [11]. This results in the modification of several physiological processes, such as heart activity, respiration, blood pressure, pupil dilation and many others [12]. For this reasons, several methodologies and instruments (obtrusive and unobtrusive) have been used in the last years in order to identify and classify physiological stress markers. According to [13], Heart Rate (HR), Heart Rate Variability (HRV) and Electrodermal Activity (EDA) represent the primary measures used for stress detection and classification. Recently, other studies have introduced also EEG, Blink Reflex through surface EMG, and eye tracking data analysis [14], but the instruments to collect these physiological signals could not be easily accepted by frail older people.

In our study we mainly refer to the detection of acute stress, related to specific training sessions, in frail MCI older adults living in a long-term care (LTC) facility, which are already affected by a chronic stress condition. We focused on the collection of HR, HRV and EDA for stress detection purposes by commercial wearable sensors and a customised m-health solution. Each participant has been monitored during specific training sessions based on standard rehabilitation activities commonly used in the daily LTC facility. Activities consist in cognitive training alternated by a light physical exercise by using a cycle-ergometer. In fact, clinicians observed that cognitive performances can be improved by the physical activity, but they have no information about how this can affect the stress conditions of frail subjects. Therefore, the proposed system is designed to highlight the short-time improvements of cognitive performances generated by the proposed physical exercise, and the related stress response.

In fact, by increasing or changing either the physical activity or the cognitive training to further stimulate the subject's abilities, the level of stress can increase, generating further risks for the frailty condition. For this reason, and for the subjective nature of the stress response, it is important to provide a novel solution for automatic stress detection. To this aim, we exploited the physiological dataset, collected during the pilot study, to evaluate the performances of a binary stress detection system, based on different machine learning (ML) algorithms. It provides promising results with respect to standard solutions conducted in controlled settings. In fact, in this case, we conducted the pilot study in a semi-controlled setting, by performing the activities in ambulatory and with the medical support, but inferring stimuli without a predefined stress response. However, this system is not suitable to infer and quantify the level of stress generated by a specific training activity, and consequently personalise the treatment. For this reason, we propose to enrich the m-health solution through the definition of a novel Decision Support System (DSS) for online stress monitoring. This solution is based on the analysis of physiological signals during the therapy in a time- or event-based manner, which results are then used as input for a decision module able to infer the user stress level and support the medical user in the definition of a personalised training activity. The complete system is thus composed by a mobile solution, aimed at collecting data and extracting relevant features by making a preliminary signal processing, and a DSS cloud-based system that provides its output to the m-health app.

A. MAIN CONTRIBUTIONS

The main contributions of this work can be summarized as follows:

- The pilot study demonstrates the efficacy of the physical activity on the cognitive performances in frail older adults, while maintaining a low stress level.
- We use the real dataset, collected in a semi-controlled environment, to evaluate the performances of a binary stress detection system based on different classification algorithms. It provides promising results compared with solutions used in completely controlled settings.
- Finally, we propose a novel Decision Support System (DSS) with a mobile pervasive architecture for online stress monitoring. This solution supports users and the medical specialists with a more valuable and detailed feedback during the therapy, in order to provide personalised treatments. Our proposal can be exploited both in clinical settings and in remote monitoring systems.

B. PAPER OUTLINE

The paper is organised as follows. Section II presents an overview of the related works and motivations in terms of: (i) the effects of physical activity on cognitive performances in older adults, and (ii) stress monitoring and detection systems. Sections III and IV present material and methods of the randomised cross-over pilot study and stress classification, respectively. Section V presents and discusses the results of both studies. Section VI describes our proposal of a DSS for online stress monitoring and the mobile system architecture. Finally, Section VII draws conclusions and future works.

II. RELATED WORKS

In order to better highlight the novelty of the proposed solutions and results, we provide a brief summary of the related works in two different research fields: (i) methodologies and results on the impact analysis of physical activity on cognitive performances in older adults, and (ii) the state of the art in stress monitoring and detection systems.

A. PHYSICAL ACTIVITY AND COGNITIVE PERFORMANCES IN OLDER ADULTS

In the last decades, several randomised controlled trials (RTCs) have been conducted to evaluate the impact of physical activity on cognitive performances and mental health in older adults [15], [16]. Most of them involve healthy subjects, able to execute long-term physical exercise programs, and assess the cognitive improvement mainly through the clinical evaluation of different cognitive functions (e.g., through neuro-psychological tests or a test battery). Among the reviewed studies, 25 of them measure attention and processing speed, 17 memory recall (immediate and/or delayed), 20 executive functions, and 13 working memory. Reported RTCs present considerable heterogeneity in terms of subjects' characteristics, exercise programs, treatment duration, samples' size, adherence rate, and cognitive tests. Nevertheless, those studies provide a general evidence of the improvements on the cognitive functions generated by the physical exercise. On the other hand, considering MCI older adults, a recent review points out that there is no general evidence of the positive effect of physical exercise on cognitive functions for this target population [17], and more rigorous and focused studies are necessary to identify possible benefits. Specifically, [18]–[20] have demonstrated improvements in one or more executive function domains by applying heterogeneous, supervised, and group-based training programs using different exercise categories (e.g., aerobic, resistance, and multi-modal). Conversely, other studies do not present any significant improvement [21], [22]. This difference can be due to both the heterogeneity of the studies and the different clinical type of MCI. In fact, MCI is currently distinguished in uni-domain and multiple-domain, based on the number of impaired cognitive functions, which can generate different reactions to physical and cognitive stimuli. For these reasons, it is important to define focused studies on homogeneous populations as preliminary investigation towards personalised treatments and training programs aimed at maintaining subjects' main abilities. Furthermore, since frailty often includes a chronic stress condition due to the management of multiple chronic illnesses, it is important to investigate if the physical-cognitive training has also a positive or negative impact on the stress condition, considering that it is a supplementary requested activity. To the best of our knowledge, there is no specific study neither on the effect of acute physical exercise on the cognitive performances of MCI frail older adults, or on the stress generated by this type of training. To this aim, we defined a specific cognitive and physical training protocol that has been executed with a group of MCI frail older adults as a randomised cross-over pilot study.

B. STRESS MONITORING AND DETECTION SYSTEMS

Classification algorithms of the physiological response to stressful and non-stressful conditions have been deeply studied in the last years [23]. HR, HRV and EDA represent the reference physiological signals for stress detection [13], and they can be accurately monitored through commercial wearable devices. Physiological markers are often matched with the analysis of cortisol levels, which represents the biological gold standard for stress detection, to generate a more accurate evaluation of the stress response [24]. However, cortisol sampling is an invasive method, it does not allow a continuous monitoring and requires complex laboratory analysis, with a significant delay in the stress detection phase, not suitable for the system we are envisioning. In other studies, EDA and HRV are also integrated with EEG [25], physical activity data based on 3D accelerometer [26], video-based activity data [27], facial expressions [28], and speech analysis [29].

Other research studies have investigated different biosignals as stress detection markers, such as pupil diameter [30], and eye gaze [31]. Thermography through thermal camera or IR touchscreen has been used to extract stress-related markers, such as breathing [32], facial blood flow changes [33], and photoplethysmographic (PPG) signal [34], [35]. Facial hyperspectral imaging (HSI) method has also been tested in order to obtain tissue oxygen saturation (StO2) for stress detection [36]. Finally, some study also used behavioural data, especially while performing computerised tasks, including keystroke dynamics [37], and typing behaviour on a smartphone using accelerometer and gyroscope data [38].

Most of these studies are conducted in laboratory settings, alternating non-stressful periods with validated stress induction tests, such as Trier Social Stress Test (TSST) [39], Stroop Color-Word Test (SCWT) [40] , Sing-a-Song Stress Test (SSST) [41], and cold pressor test [42]. Moreover, the stressing protocols are designed in such a way that all the stressors' characteristics (i.e., intensity, duration, timing, frequency) are rigorously controlled. Therefore, the system has an implicit ground truth in the classification of stressful and non-stressful events.

On the other hand, a limited number of studies attempted to move from laboratory settings, performing stress detection

in work offices [43], [44], university campus [45], [46], automobile environments [47], [48], and finally targeting unrestricted daily life conditions [49], [50]. These studies use self-reports as stress ground truth, which are collected by using self-report questionnaires and scales, such as Cohen Perceived Stress Scale (CPSS) [51], and Ecological Momentary Assessment (EMA) [52]. However, self-report measurement suffers from several forms of recall bias [53], which can be alleviated by prompting users to report their status multiple times in their current environment. Nevertheless, participant burden is a major drawback in this approach, due to the repeated administration of multi-item questionnaires.

Moreover, stress detection studies don't usually focus on specific target populations, and they mainly involve young healthy subjects, whereas the number of studies targeting older adults is really limited. To the best of our knowledge, stress in older adults is usually assessed using standard self-report questionnaires or clinical rating scales. [54] is the only study we found aimed at detecting stressful and non-stressful events in activity of daily living (ADL) of 6 older subjects with dementia, and it is based only on EDA monitoring. In this case the authors used the care giver subjective notes as ground truth on the stress condition of each subject, and they applied predefined signal thresholds to determine both generalized and personalized stress levels, without exploiting any data learning or modeling technique. No additional stress detection studies using physiological data and sensor technologies with older adults are reported. A particular work has been presented in [55] where a step watch is used to measure the wandering behavior of older people suffering from dementia. The authors exploit this activity as individual stress indicator, even though there is no generalised assumption on this.

In our reference scenario we are focusing on MCI frail older adults, characterised by a chronic stress condition, monitored during specific rehabilitation and training sessions, which represent part of the general activity conducted in a LTC facility. Despite the ambulatory environment where the daily training and rehabilitation activities are usually carried out in a LTC facility may be considered similar to a laboratory, our reference scenario represents an intermediate condition between lab and natural settings (i.e., "into-the-wild".) In fact, we can define our stress induction protocol as semi-controlled considering the following conditions:

- The cognitive stressor (SCWT) has not a predefined duration and intensity, since it depends on individual cognitive abilities.
- There is no evidence found in the literature that the proposed physical activity, suitable for our target population, can be considered an intended stressors. We just set the duration of this physical exercise, and we investigated its effect on both the physiological stress response and cognitive performances.

We refer to HR, HRV, and EDA as physiological stress markers, and we analysed in the literature the reference learning algorithms used for stress detection. Stress detection is usually defined as a supervised learning problem, and Support Vector Machines (SVMs), Decision Trees (DTs), Random Forests (RFs), k-Nearest Neighbours (k-NNs), Bayesian Networks (BNs), AdaBoost (AB), and artificial neural networks (ANNs) are the most widely used learning schemes in this research field [56]. Alternative solutions are based on Fuzzy Inference Systems (FISs) [57], [58] or clustering techniques [59]. In the first case, FISs are mainly characterised by predefined Membership Functions (MFs) and rules, whereas clustering use iterative approaches to determine the optimal number of clusters. These solutions suffer from the main drawback of knowledge elicitation for system modeling and learning from collected data. Moreover, unsupervised approaches based on Self-Organizing Maps (SOMs) have been used, thus avoiding the collection of stress class labels [60], [61]. However, they require an additional effort to analyse the patterns emerged onto the map, and a comparison with a clinical stress evaluation is still required as a system validation.

III. THE PILOT STUDY

In this section we present in detail the materials and methods used in the pilot study design, and in the evaluation of stress detection algorithms based on different ML techniques.

A. PARTICIPANTS

We conducted a randomised cross-over observational study in collaboration with LTC facility iCARE, located in Viareggio (Lucca, Italy)[1], which is interested in improving its assistance services by implementing personalised treatments. The study received the Ethical Clearance both by CNR Ethical Committee[2] and by iCARE internal committee. All the participants have been deeply informed about the study methodology, specifying the purpose, the technology functionality, the monitoring sessions and the data management phases, and they signed the informed consent to participate in the study.

Recruitment of older people, especially those suffering from multimorbidity and frailty , is a major challenge for most research studies struggling to find participants among these population groups [62]. Figure 1 shows the CONSORT[3] Flow Diagram for transparent reporting of trials. 57 subjects have been assessed for eligibility in the nursing home; 24 did not provide the informed consent to participate in the study. Among those who signed the informed consent, 20 did not meet the inclusion criteria reported in Table I. In addition, 4 subjects dropped out before the study began (i.e., some of them moved to another LTC facility, and others had health complications). Eventually, we enrolled 9 frail older adults, 5 women (mean age 83.6 ± 4.3 yo) and 4 men (mean age 74.7 ± 11.6 yo).

Psycho-physical and behavioural characterisation of the study population has been conducted before the beginning

[1] http://www.rsatabarracci.it/

[2] CNR protocol number 0060896/2017 date Sept. 25, 2017.

[3] http://www.consort-statement.org/

of the study, based on the evaluation of the following clinical rating scales:

- Mini Nutritional Assessment (MNA) for a screening of the nutritional status.
- Activity Daily Living (ADL) and Instrumental Activity Daily Living (IADL) for functional autonomy.
- Apathy Evaluation Scale (AES) and Hamilton Depression scale (HDS) for mood and behaviour.
- Cohen Perceived Stress (CPS) for stress awareness.
- Insomnia Severity Index (ISI) for sleep quality.

The obtained scores, shown in Table II, classify most of the subjects in a normal nutritional status (MNA $\geq$ 23.5), in a low index of dependence in ADL (ADL $\geq$ 5), with a severe to moderate dependence in IADL, which is mainly caused by the lack of practice in specific activities due to living in a nursing home (3 $\leq$ IADL $\leq$ 5). The remaining scores highlight that most of the participants presents mood disorders (moderate to severe depression conditions, HDS $\geq$ 18), 33% presents reduced initiative and apathy (AES $>$ 37.5) and 33% reports high level of perceived stress (CPS $\geq$ 20). The entire group presents light (8 $\leq$ ISI $\leq$ 14) to moderate (15 $\leq$ ISI $\leq$ 21) insomnia. Moreover, we evaluated all the subjects through a brief pathological and pharmacological assessment, from which emerged the presence of multiple chronic diseases under personalised pharmacological treatment for the entire group.

Screening of cognitive performances has been carried out through Mini Mental State Examination (MMSE). MMSE provides a clinical picture compatible with the diagnosis of MCI for all the subjects, with particular reference to attention, memory recall, and executive functions, which is further supported by the partial functional autonomy highlighted by IADL. Then, each subject executed an additional neuropsychological test battery, composed by Frontal Assessment Battery (FAB), Trail Making Test (TMT A, B, B-A), and Disyllabic Word Span Test (DWST) to specifically assess these cognitive domains. MMSE scores, together with neuropsychological scores, are reported in Table III. 6 subjects exhibit reduced performance in executive functions (FAB $<$ 13.5), 3 subjects present attention deficits (TMT-A $>$ 94, TMT-B $>$ 283), whereas almost all present reduced memory recall (DWST equivalent $\leq$ 3). Subject #5 refused to complete TMT and, for this reason, the execution time for both TMT-B and TMT-B-A is not assigned (*N.A*). Therefore, neuro-psychological evaluation supports the diagnosis of multi-domain MCI, with reduced performances in all the analyzed cognitive domains.

B. INSTRUMENTS

We designed and developed the m-health solution used to monitor and evaluate the relationship between cognitive performances and physiological stress response. The monitor functionalities have been briefly presented in [63], whereas in this paper we present its application on a real group of subjects and the evolution towards stress detection and DSS.

The system consists of two different wearable sensors, aimed at collecting HR, HRV and EDA, connected to a mobile app in charge of sensors' synchronisation, data streaming and storage, both on the mobile device and on a remote server used for signal processing and analysis. In order to guarantee a good user acceptance of the system, we selected two commercial mobile devices: Zephyr BioHarness3[4] (Figure 2a) for ECG monitoring, and Shimmer3 GSR+ Development Kit[5] (Figure 2b) for EDA.

ECG signal is recorded with 250 Hz sampling rate, since a sampling rate lower than 200 Hz may create a jitter in the QRS complex recognition, thus introducing errors in the reconstruction of RR intervals. Moreover, a lower sampling rate can generate an error in the HRV spectrum estimation, which increases with frequency [64]. The chosen sampling rate also allows to save both bandwidth and storage space compared to higher rates, such as 1 KHz. ECG is processed on board to extract HR and HRV. HR is reported at 1 Hz, whereas HRV is derived from the detection of the time interval between consecutive R-peaks in the ECG signal.

Shimmer3 GSR+ consists of a wrist-worn unit and 2 electrodes to be applied to fingers of the non-dominant hand, and it provides both skin resistance (KΩ) and conductance (μS) values with 51.2 Hz sampling rate.

The mobile app is developed for Android OS. As a first step, it sequentially connects both devices through standard Bluetooth (BT) or Bluetooth Low Energy (BLE) interfaces. Then, once the medical specialist selects the specific monitoring protocol, the app performs the necessary setup, such as synchronising the devices' real time clock (RTC) with the smartphone, selectively enabling and/or disabling sensors, and setting different sampling rates. It is also in charge of starting the data stream from the wearable devices and parsing, processing and storing all the physiological data during the experiments. The data stream can be also displayed for real-time data visualization.

In terms of protocol execution, the mobile app allows the medical specialist to switch between the successive phases (as detailed in Figure 3), so that the system is able to precisely match each signal time window to the corresponding protocol phase.

C. EXPERIMENTAL SETUP

The experimental protocol is designed to analyse the impact of physical activity both on the cognitive performances and the related physiological stress response of MCI frail older people. To this aim, in our study design we asked participants to execute the Stroop Color-Word Test (SCWT) as cognitive training, alternate to a light physical exercise or a rest period. The physical exercise (E) and rest (R) are considered as conditions in the randomised cross-over study, in which each subject is allocated for both interventions. In terms of physical activity, we decided to use a cycle-

[4]www.zephyranywhere.com
[5]www.shimmersensing.com

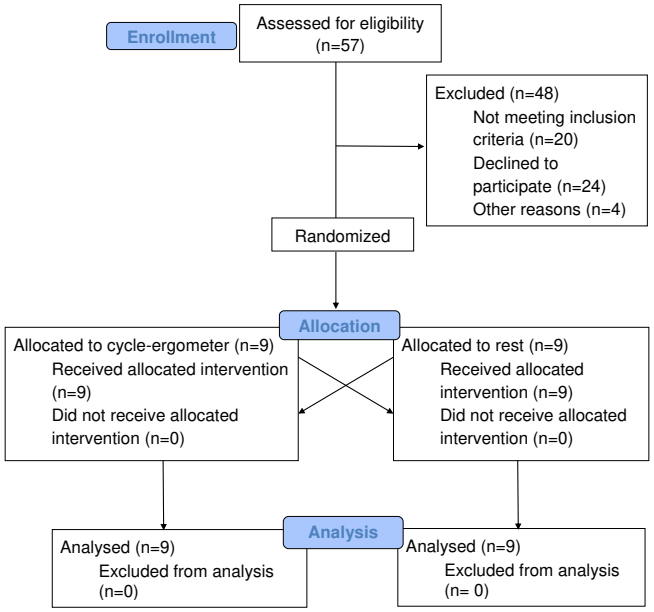


FIGURE 1. CONSORT Flow Diagram of the randomised cross-over study

TABLE I. Inclusion and exclusion criteria for the randomised cross-over pilot study

Inclusion criteria	Exclusion criteria
Age $\geq$ 65 yo, both genders	Diagnosis of dementia (MMSE $<$ 23)
Living in the nursing home for at least 6 months	Any already diagnosed psychiatric pathology
No absolute contraindications to perform aerobic exercise	Pacemaker or any implantable electrical device

TABLE II. Clinical rating scale scores.

SUBJECT	*MNA*	*ADL*	*IADL*	*AES*	*HDS*	*CPS*	*ISI*
#1	27	6	3	59	34	18	15
#2	28	6	5	37	45	21	21
#3	26	6	4	9	7	12	10
#4	26	6	3	40	19	20	18
#5	25	6	3	52	28	11	11
#6	26	6	3	35	22	17	22
#7	23	6	3	32	25	22	18
#8	25	5	5	11	7	12	15
#9	25	6	3	18	6	11	13

TABLE III. Neuro-psychological test scores. DWST scores are corrected according to age and education level of each subject.

SUBJECT	*MMSE*	*FAB*	*TMT-A*	*TMT-B*	*TMT-B-A*	DWST	DWST equivalent
#1	25.2	13.7	100	288	188	3.25	1
#2	27.2	15.7	26	62	38	3.75	2
#3	28.3	8.3	24	49	25	3.25	1
#4	26.4	15.7	27	106	79	3.50	2
#5	25.4	11.2	117	*N.A*	*N.A*	3.75	2
#6	25.4	9	112	198	86	4.25	3
#7	23.1	12	30	102	72	3.75	2
#8	23	13	89	140	51	3.50	2
#9	28	13.4	49	76	27	4	3

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access

(a) Zephyr Bioharness 3 (b) Shimmer3 GSR+ Unit

FIGURE 2. Wearable devices.

ergometer, as frequently used in clinical trials, especially with this subject category. We set the length of the exercise to 12 minutes, considering that clinical evidences have shown that physiological parameters change related to the physical activity can be observed already after 6 minutes of cycle ergometer training [65].

SCWT is a validated, quickly and non-invasive mental stress test, widely used in the literature to assess the inhibitory control. Inhibitory control is one of the core components of the executive functions, and it involves the control of attention, behaviour and emotions to override a dominant behavioural response and select a more appropriate behavior that is consistent with a specific goal [7]. We used a short-form of the standard SCWT composed by 3 consecutive phases:

1) Color reading phase: color names are written with black ink, and the subject should simply read them in the order they are presented.
2) Color recognition phase: there are colored patches (with the same colors presented in the previous phase), and the subject should say the color of the patches in the order they are presented. Phase 1 and 2 constitute the congruent part of the SCWT.
3) Color interference phase: color names are written with different colored inks, and the subject should say the ink color of each color name. This phase represents the incongruent part of SCWT, demanding a stronger cognitive effort due to the inhibitory control.

In the study protocol, each subject performs the SCWT before and after the condition (i.e., R or E). For each test, we analyze the results in terms of execution time ($Time_{ex}$), and accuracy (i.e., number of errors, $Errors$), for each test phase and for the whole test. Execution time and accuracy of each test phase are then combined to obtain two different scores, which represent respectively the $Time_{effect}$ and $Interference_{effect}$ for the overall test. Both scores are also corrected on the basis of age and education level of each subject [66].

As shown in Figure 3, each session consists in the following phases:

- Baseline (B), in which the subject sits on a chair in a rest condition for 3 minutes.
- 1st SCWT (S_1).
- 3-minute recovery (Rec_1).
- 12-minute exercise (E) or rest (R) condition.
- 3-minute recovery (Rec_2), made in both cases to keep the same protocol design.
- 2nd SCWT (S_2).
- 3-minute recovery (Rec_3) before stopping the session.

In order to further engage the study participants, the LTC facility provided us with a dedicated ambulatory room, designed for the monitoring sessions, in which the medical specialist receives the subjects and help him/her wearing the sensors before the session starts. At this moment, the m-health app connects to the devices, performs all the necessary setup and finally begins the data streaming. Each subject performed the two sessions (Rest and Exercise) in a random order with a minimum distance of 1 week. As shown in Figure 1, all the enrolled subjects successfully executed both interventions.

IV. METHODS

A. PHYSIOLOGICAL DATA ANALYSIS

In this section we present the main methods used to analyse the physiological signals collected during the training sessions.

1) Data pre-processing and feature extraction

After checking the integrity of the collected data due to the wireless communication between the wearable devices and the smartphone, we evaluated the reliability of the collected samples in the signal time series. As far as ECG is concerned, R-peak detections are computed from contiguous 250 ms blocks of raw ECG data, which are processed on board to account for false or missed R-peak detections. HR is computed mainly from the preceding 15 seconds of ECG data. Zephyr BioHarness3 exploits an on-board worn detection algorithm to indicate HR data collected when the chest strap is not correctly worn. Moreover, signal-noise ratio (SNR) of ECG signal is also used to identify extremely noisy ECG samples, which allows to establish HR confidence and indicate reliability of each 1-second HR sample. As far as HRV is concerned, we considered values in the range $[250 - 2400]$ ms, as suggested by the vendor, which were not time-aligned with unreliable HR samples. All unreliable samples were considered as missing values in the corresponding signal time series. Finally, HR and HRV time series consisted in $\geq 95\%$ of reliable samples.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access

REST	B	S_1	Rec_1	R	Rec_2	S_2	Rec_3
	3'		3'	12'	3'		3'
EXERCISE	B	S_1	Rec_1	E	Rec_2	S_2	Rec_3

FIGURE 3. Sequence and duration (minutes) of the different phases of the cross-over study design. SCWT (S_*) has not a fixed execution time since it depends on the subjects' ability.

TABLE IV. Features extracted from HR, HRV, and EDA.
(*bpm* = beats per minute. *s* = seconds. *n.u.* = normalised units. *S* = Siemens)

Feature Name	Signal	Description
max	HR/HRV	Maximum value of HR/RR intervals (bpm/ms)
min	HR/HRV	Minimum value of HR/RR intervals (bpm/ms)
max-min	HR/HRV	Difference between maximum and minimum value of HR/RR intervals (bpm/ms)
mean	HR/HRV	Mean value of HR/RR intervals (bpm/ms)
median	HR/HRV	Median value of HR/RR intervals (bpm/ms)
std	HR/HRV	Standard deviation of HR/RR intervals (bpm/ms)
RMSSD	HRV (time domain)	Root Mean Square of the Successive Differences between RR intervals (ms)
NN50	HRV (time domain)	Number of adjacent RR intervals with an absolute difference $\geq$ 50ms;
pNN50	HRV (time domain)	% of NN50 within the analysed time window
Tot. spectrum power	HRV (frequency domain)	Total spectrum power (ms^2)
VLF	HRV (frequency domain)	Very Low Frequency ($0.003-0.04$ Hz) power band (ms^2)
LF	HRV (frequency domain)	Low Frequency ($0.04-0.15$ Hz) power band (ms^2)
HF	HRV (frequency domain)	High Frequency ($0.15-0.40$ Hz) power band (ms^2)
LF norm	HRV (frequency domain)	LF divided by the total spectrum power minus VLF band (n.u.)
HF norm	HRV (frequency domain)	HF divided by the total spectrum power minus VLF band (n.u.)
LF/ HF	HRV (frequency domain)	Ratio between LF and HF power bands (n.u.)
nSCRs	EDA	Number of above-threshold Skin Conductance Responses
Cum Amps	EDA	Cumulative amplitudes of all the detected SCRs (μS)
ISCR	EDA	Area (i.e., time integral) of the all the detected SCRs (μS·s)
mean SCL	EDA	Mean Skin Conductance Level value (μS)
mean EDA	EDA	Mean value of the overall EDA signal (μS)
max EDA deflection	EDA	Maximum positive deflection of the overall EDA signal (μS)

Then, we divided the physiological signals into their respective protocol phases, by using the start and end timestamps of each phase recorded by the application. We use Matlab (v.R2017b) for features extraction from HR and HRV, whereas EDA is processed by using the Matlab-based software Ledalab (v3.4.9)[6]. Table IV presents all the features extracted from HR, HRV (both time and frequency domains) and EDA.

For HRV spectral analysis, we estimated the power spectral density (PSD) by applying the Lomb-Scargle algorithm [67]. It provides the information about how the power of HRV is distributed with respect to the frequency. This approach let us to work directly with HRV as a not uniformly sampled signal, thus avoiding an initial interpolation to convert HRV in a signal with a fixed sampling rate, usually set between 2 Hz and 5 Hz. In this way, we avoid any additional artifacts that can be generated by adding new harmonic components to the original signal.

As far as EDA processing is concerned, we applied a Butterworth 1^{st}-order low-pass filter with 5 Hz cut-off frequency to remove the high frequency noise, since the signal dynamic is usually considered to be in the range $[0-5]$ Hz [68]. Motion artifacts, especially during the physical activity, could be considered negligible, since each participant performed the whole experimental protocol in a sitting position (included cycle-ergometer training), with the wearable devices attached on the upper part of the body. However, they have been completely removed through low-pass filtering. Then, we performed a Continuous Deconvolution Analysis (CDA) in order to decompose EDA into its Skin Conductance Response (SCR) and Skin Conductance Level (SCL) components [69]. SCL (i.e., tonic component) is the slowly changing part, and it is usually considered as the baseline portion of the signal, whereas SCR (i.e., phasic component) is the faster changing part, and its behavior can be related to specific or not specific stimuli. After CDA, we defined a response window of the same length of protocol phase under investigation, and a minimum signal amplitude threshold of 0.01μS, in order to detect all the above-threshold SCRs.

The feature extraction process resulted in 10 features vectors

[6]http://http://www.ledalab.de/

for each session, including 4 features vectors for both S_1 and S_2 (i.e., the 3 test phases and the whole test), baseline (B) and condition (R or E) phases. Each feature vector is composed by 28 physiological features, and the outcomes of the SCWT for the cognitive tests.

2) Statistical Analysis

Each subject performed the Rest and Exercise sessions in a random order. Therefore, in order to assess the independence of all the sessions, we examined separately those executed in the first cycle and in the second one by evaluating any significant difference between B and S_1 (full test) phases of each cycle. Then, we compared the physiological features and cognitive outcomes obtained in S_1 and S_2, in the Rest and Exercise sessions, separately. At the same time, we also evaluated any significant difference in S_2 among Rest and Exercise sessions. This analysis provides an overview of the impact of the physical activity on both cognitive performances and the physiological stress response. We also compared physiological features during R and E conditions to assess the physiological impact of the proposed physical activity. Then, in order to evaluate the adaptive stress response throughout the monitoring protocol, we evaluated any significant difference in the physiological features related to B and S_2 of both sessions.

Regarding the statistical analysis, we used Shapiro-Wilk to test the normal distribution of the data, and Levene's test to assess the homogeneity of variances. Paired t-test is used to make the comparison, in case both normality and homoscedasticity assumptions are met, Wilcoxon signed-rank test otherwise. We performed one-side test for both tail sides ($\alpha = 0.05$).

B. STRESS DETECTION

After the the statistical analysis, we investigated the performances of several ML techniques for stress detection.

1) Dataset Preparation

For each monitoring session, we considered the following protocol's phases: baseline (B), SCWT (both S_1 and S_2) and condition phases (E or R). Recovery periods (Rec_1, Rec_2, and Rec_3) have been excluded from analysis, since we cannot estimate the time necessary for each subject to return to the initial physiological condition after stressor application. For this reasons, recovery periods may have an intermediate stress nature [25]. We divided each phase into 1-minute consecutive time windows with 50% overlap to maximize the number of available frames and to allow the extraction of meaningful stress-related HRV features even in short-time windows [70]. Since SCWT duration is not fixed a priori, but it depends on the individual cognitive abilities, we considered the full test duration if the cognitive test lasted less than 3 minutes, otherwise we selected the last 3 minutes for both S_1 and S_2. In our study only 2 participants have been able to execute at least one of the 2 cognitive tests in less than 3 minutes. This let us select similar time windows for S_1 and S_2 and to include the last (i.e., incongruent) phase of the SCWT, which usually requires the greater mental effort and related stress response. This approach resulted in 38 instances for each subject, including 5 instances for B, S_1 and S_2, and 23 instances for R and E. For each instance we extracted 28 features from HR, HRV, and EDA, which are listed in Table IV.

Then, we manually labeled each dataset. As far as rest and cognitive activities is concerned, we used the study protocol as an implicit binary stress ground truth, assuming that the SCWT generates a stress condition as extensively shown in the literature [71]. For this reason, S_1 and S_2 instances have been labeled as *'Stress'* (*'S'*), whereas B and R instances as *'No Stress'* (*'NS'*). On the other hand, since we are experimenting a particular type of physical activity that cannot be considered as a validated stressing agent, no preliminary assumptions can be made. Thus, we rely on the statistical results shown in V-A3 assuming that the proposed physical activity is non-stressful (*'NS'*). We refer as R^* to the dataset composed of the Rest sessions of all subjects, whereas E_{NS} represents the dataset for Exercise sessions. Each dataset is composed by 342 instances.

2) Model Selection and evaluation

We investigated the performance of some stress detection algorithms, based on a set of ML classifiers. Specifically, we used BN, SVM, k-NN, C4.5 Decision Tree (J48 implementation), RF, and AB learning methods. Regarding AB, we selected decision trees as weak learners. Model selection was performed through 10-time stratified 10-fold cross-validation. We performed hyperparameter tuning for each classifier by applying an exhaustive grid search over a subset of the hyperparameter space of each learning algorithm, selecting the best performing configuration. We used WEKA[7] ML engine to train our learning schemes.

In our scenario, it is fundamental to keep in mind that all the datasets are unbalanced. The minor class (*'S'*) represents the 26.3% of the whole dataset for both R^* and E_{NS} dataset, with a ratio between major and minor class of 2.8.

For this reason, we applied stratification to preserve the class ratio within each fold, avoiding "unlucky" splittings composed by almost single-class examples. When dealing with unbalanced datasets, accuracy is not reliable as a single performance metric. In this case, training a classifier with an unbalanced dataset could create a biased model that always classify new instances with the major class. This can still provide a good accuracy ($\sim 74\%$ for a biased classifier in our case), which increases as the imbalance increases. Considering minor class instances as positive instances, precision and recall (i.e., sensitivity) represent more reliable metrics since they are not affected by the imbalance, as they focus only on the ability of detecting positive instances. For these reasons, accuracy should be evaluated together with precision and recall. Moreover, despite the majority of studies involving bi-

[7]https://www.cs.waikato.ac.nz/ml/weka/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access

nary classifiers with unbalanced datasets use the Area Under Receiver Operating Characteristic (AUROC) curve as main evaluation metric, it has been demonstrated that the Area Under Precision Recall Curve (AUPRC) represents one of the most fair, informative, and powerful metrics for unbalanced cases [72]. According to these considerations, we referred to precision, recall, and AUPRC, by selecting classifiers which exhibited the best trade-off among these metrics.

Afterwards, for each dataset we applied Synthetic Minority Oversampling TEchnique (SMOTE) on training data to balance the class ratio in the training folds [73]. This let us to compare the classification performances by using both original unbalanced training sets, and synthetically-balanced ones, testing the models only on real data, and obtaining thus more realistic performances. Regarding the balanced datasets, we selected learning schemes which exhibited the best accuracy.

Although ML analysis has been performed in offline settings, we also provide a preliminary evaluation of the training time of the best-performing configuration of each classifier. For this evaluation, we used the algorithm implementation provided by the WEKA ML engine, using the following settings:

- Notebook Dell XPS 9343;
- Windows 10 OS;
- Intel(R) Core(TM) i5-5200U CPU @ 2.20 GHz;
- 8.0 GB RAM;

This analysis has been performed for both R^* and E_{NS} datasets, since the best-performing model might be different among them. Results are averaged among 10 repetitions of the 10-fold cross validation scheme, in order to take also into account changes induced by the creation of different folds and by the order in which single instances are processed.

3) Feature Selection and Reduction

All the classifiers have been initially trained using all the extracted physiological features as input data. Then, for each dataset, the best performing classifiers have been further trained using a subset of the most relevant features, which have been automatically selected within the training set. Specifically, we applied:

1) Correlation-based feature selection (CFS) [74];
2) Information gain ratio-based feature selection [75];
3) Principal Component Analysis (PCA) as feature reduction technique [76]. We set variance threshold to 95%.

Training time has also been evaluated for these cases.

V. RESULTS AND DISCUSSION

In this section we present and discuss the results of the pilot study in terms of statistical analysis of the collected dataset, and the performances of the ML algorithms for stress detection.

A. PILOT STUDY

We divided the results based on the different objectives of the study and, first of all, we present the assessment of the sessions' independence, as a preliminary condition for the following analysis. To this aim, we evaluated any significant difference between B and S_1 (full test) of all the sessions as belonging to two separate groups: the first cycle and the second cycle. Significant differences with their corresponding p-value are reported in Table V. In both cycles, mean EDA, ISCR and HRV NN50 are significantly greater in S_1 than in B, whereas HRV pNN50, nSCRs and SCRs cumulative amplitude present the same behaviour only in the second cycle.

Results presented in Table V show quite high standard deviation values in both cycles, which may be mainly due to the small sample size. However, a similar trend appears in both cycles, witnessing an increased physiological response to the cognitive workload with respect to the baseline. Moreover, the increasing trend shown only in the second cycle is in contrast with a possible habituation and relaxation effect that could emerge in repeating the protocol. Therefore, we can assume that the physiological response at the first step of the two sessions is similar and we can consider each session as independent from the other for subsequent analysis.

1) Cognitive tests comparison

In order to highlight the effect of the physical activity on the cognitive performances of MCI frail older adults, and on their physiological stress response, we investigate the significant differences among the analysed cognitive outcomes and the physiological features obtained in both Rest and Exercise sessions, separately. Results are presented in Table VI, divided by the corresponding test phase. As far as the cognitive outcomes in Rest sessions is concerned, the subjects spend significantly less time to perform S_2 than S_1, while introducing a significantly higher number of errors in S_2. The $Time_{effect}$ and $Interference_{effect}$ correct scores further highlight this behaviour, since the former decreases in S_2, whereas the latter increases. By comparing the single phases of each test, we notice a higher number of $Errors$ and a lower execution time ($Time_{ex}$) for S_2 phase 3 (i.e., SWCT incongruent phase).

Instead, in the Exercise sessions, the subjects spend less time to complete S_2 (especially in phase 3), and $Errors$ are reduced in all S_2 phases, significantly improving the overall cognitive performance with respect to S_1. By comparing the cognitive performances in S_2, between Rest and Exercise sessions, we notice also that $Errors$ is significantly lower in the Exercise sessions than in the Rest ones, especially in phases 2 and 3. In addition, $Time_{ex}$ is reduced in the Exercise sessions in phase 1, in which all the subjects generally obtained the maximum score. Finally, in terms of full test, the $Interference_{effect}$ correct scores are lower in the Exercise session than in the Rest ones.

As far as the physiological features is concerned, by analysing the Rest sessions we notice that the mean EDA value is significantly higher in S_2 than in $S1$ (full test and all phases), the same for the cumulative amplitude of the detected above-threshold SCRs (full test), ISCR (phase 1),

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access

TABLE V. Significant differences observed between baseline (B) and S_1 for both sessions ($\alpha = 0.05$). Feature values are reported as mean $\pm$ standard deviation

Feature Name	*Session's Cycle*	B	S_1	*p-value*
NN50	1	36.33±52.42	64.00±71.15	0.0039
	2	27.11±30.06	63.78±82.98	0.0078
mean EDA	1	1.18±2.30	1.54±3.18	0.0019
	2	1.11±2.31	1.57±2.32	0.0019
ISCR	1	0.74±1.02	1.92±2.37	0.0273
	2	0.85±1.24	1.82±1.95	0.0137
pNN50	2	12.76±15.92	15.73±18.66	0.0140
nSCRs	2	22.44±55.45	116.44±188.98	0.0156
Cum Amps	2	0.31±0.74	1.57±2.71	0.0156

TABLE VI. Significant differences observed between S_1 and S_2 in both sessions, and in S_2 between Rest and Exercise sessions ($\alpha = 0.05$).

S_{1C} vs S_{2C}			S_{1E} vs S_{2E}		
Feature Name	*Test phase*	*p-value*	*Feature Name*	*Test phase*	*p-value*
$Time_{ex}$	Full test	0.0472	$Errors$	Full test	0.0430
	3	0.0371	$Interference_{effect}$	Full test	0.0468
$Errors$	Full test	0.0104	$Time_{ex}$	3	0.0371
	3	0.0062	mean EDA	1	0.0098
$Time_{effect}$	Full test	0.0150		2	0.0273
$Interference_{effect}$	Full test	0.0066	max EDA deflection	3	0.0273
mean EDA	Full test	0.0019	max HR	Full test	0.0142
	1	0.0059		2	0.0129
	2	0.0039		3	0.0123
	3	0.0039	mean HR	Full test	0.0292
Cum Amps	Full test	0.0391		2	0.0119
ISCR	1	0.0371		3	0.0324
max EDA deflection	3	0.0059	median HR	Full test	0.0297
S_{2E} vs S_{2C}				2	0.0371
$Errors$	Full test	0.0003		3	0.0428
	2	0.0156	max HRV	2	0.0195
	3	0.0009	mean HRV	Full test	0.0226
$Interference_{effect}$	Full test	0.0019		3	0.0162
$Time_{ex}$	1	0.0143	min HRV	Full test	0.0273
				3	0.0019

and EDA maximum positive deflection (phase 3). These results show that there is a greater physiological response in repeating the SWCT test after the rest condition, which is carried out by an increasing trend in several EDA features. On the other hand, there is no significant difference in HR and HRV features between S_1 and S_2.

This trend is different in the Exercise sessions. Even though there is still a physiological response in terms of EDA features, it is mitigated in the phase 3 of S_2 presenting mean EDA values significantly lower than in the other phases, while EDA maximum positive deflection is significantly higher in phase 3. The mitigation effect is also witnessed by the SCRs cumulative amplitudes, which do not significantly differ among the cognitive tests. In addition, in terms of HR and HRV features, maximum, mean, and median HR values are significantly lower in S_2 than in S_1, and in particular for phases 2 and 3. Consequently, maximum, mean, and minimum HRV values are significantly higher in S_2, for the full test and in particular in phase 3.

From an overall point of view, these findings suggest a relaxation trend in the cardiac response in the repetition of the cognitive test after the acute physical exercise. This can also represent an indication of an adaptive stress response. In addition, there is an evident improvement in the cognitive performances.

2) Stress adaptive response

In order to investigate a possible adaptive response of the subjects to the stress induced by physical activity and cognitive training, we compare the physiological features measured in S_2 with those of B for both Rest and Exercise sessions. Table VII shows the features that present significant differences. In both cases, nSCRs, cumulative amplitudes, ISCR, mean EDA, and EDA maximum deflection are significantly higher in S_2 than in B. In the Rest sessions, HR and HRV features do not exhibit any significant differences between S_2 and B, whereas in the Exercise sessions their trend provides further support to the relaxation previously pointed out by comparing S_1 and S_2. In fact, maximum, mean, and median HR exhibit lower values in the whole S_2 even if compared with B, starting from test phase 2. Accordingly, mean HRV is higher in the S_2 full test, and in particular in phases 2 and 3, whereas minimum HRV is higher in S_2 phase 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access

TABLE VII. Significant differences observed between B and S_2 in both sessions ($\alpha = 0.05$).

B_C vs S_{2C}			B_E vs S_{2E}		
Feature Name	*Test phase*	*p-value*	*Feature Name*	*Test phase*	*p-value*
nSCRs	Full test	0.0078	nSCRs	Full test	0.0156
Cum Amps	Full test	0.0078	Cum Amps	Full test	0.0156
ISCR	Full test	0.0019	ISCR	Full test	0.0098
mean EDA	Full test	0.0019	mean ED	Full test	0.0059
max EDA deflection	Full test	0.0137	max EDA deflection	Full test	0.0195
			max HR	Full test	0.0429
				1	0.0183
				2	0.0039
				3	0.0039
			mean HR	Full test	0.0387
				2	0.0285
				3	0.0273
			median HR	Full test	0.0381
				2	0.0391
				3	0.0381
			mean HRV	Full test	0.0140
				2	0.0398
				3	0.0090
			min HRV	3	0.0102

3) Exercise and Rest conditions

Finally, we also analyse the physiological data obtained during the two condition phases: 12-minute physical exercise and rest. Results show that the physical activity impacts only on few HR/HRV features. Specifically, max HR (*p=0.0289*), min HR (*p=0.0041*), mean HR (*p=0.0022*), and median HR (*p=0.0021*) are significantly higher during the exercise phase than in the rest phase. Accordingly, mean HRV is significantly higher during rest (*p=0.0076*). These outcomes suggest that the proposed physical exercise has not a significant impact on the individual stress responses, except for the physiological increase of HR. No other features change significantly between R and E conditions, including all EDA features. This may indicate that the proposed physical activity (type, duration, and intensity) does not elicit a robust stress response.

B. STRESS DETECTION

1) Model Validation

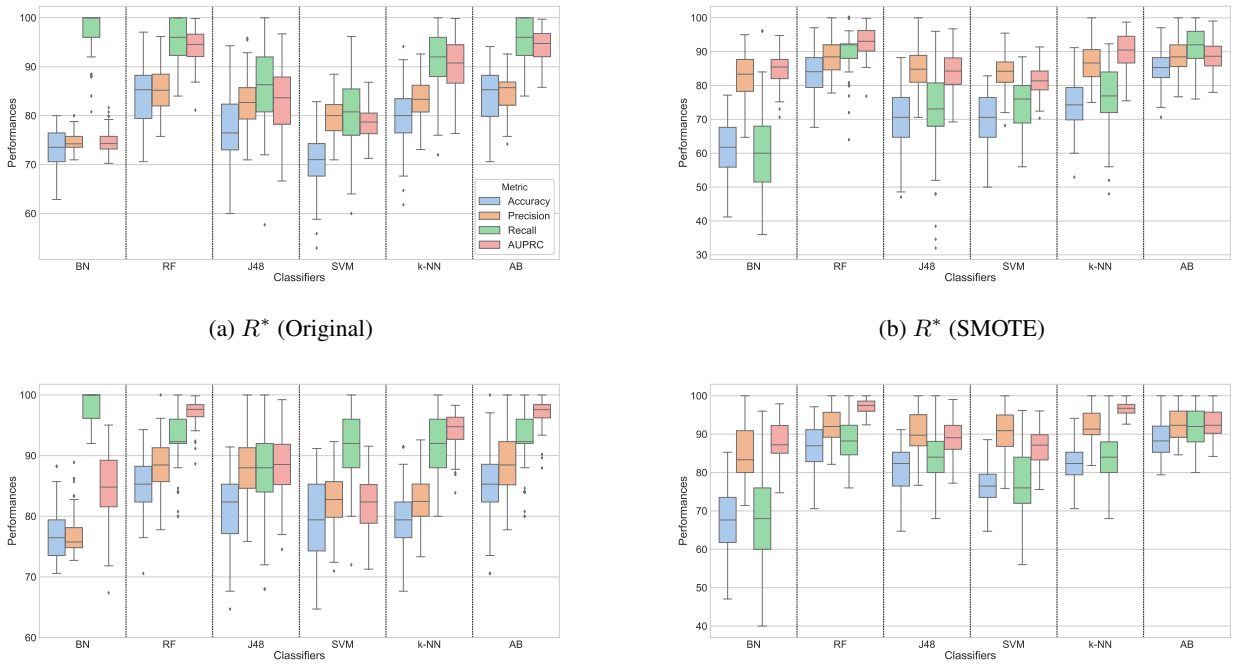
Classification performances are presented in Figure 4, divided by the reference datasets. AUPRC is reported as percentage for visualization purposes only. Random Forest (RF) and AdaBoost (AB) learning schemes outperform the other classifiers for both R^* and E_{NS}. Specifically, in case of original R^* (Figure 4a), RF presents 85.3% accuracy, 85.1% precision, 96.0% recall and 0.94 AUPRC. AB obtains similar performances with 85.3% accuracy, 85.7% precision, 96.0% recall and 0.95 AUPRC. Results on balanced R^* are shown in Figure 4b, showing again RF and AB as the best performers. Results obtained on E_{NS} dataset support our preliminary hypothesis regarding the non-stressful effects of the proposed physical activity. RF and AB still result to be the best performers with the following values: RF (accuracy = 85.4%; precision = 88.5%; recall= 92.3%; AUPRC= 0.98); AB (accuracy = 85.3%; precision = 88.6%; recall = 92.1%; AUPRC = 0.97). Results are shown in Figure 4c. Performances on the balanced dataset are in line with those obtained on the original dataset, with RF (accuracy = 87.0%; precision = 92.4%; recall= 88.2%; AUPRC = 0.97) and AB (accuracy = 88.2%; precision = 92.3%; recall= 92.0%; AUPRC = 0.92) algorithms outperforming the other classifiers (Figure 4d).

These results are related to the nature of RF and AB algorithms. In fact, they are both ensemble learning methods, which combine multiple models to improve predictions in a different manner, i.e., by using bagging and boosting methods. Specifically, RF creates a large number of relatively uncorrelated individual decision trees, and it trains each decision tree on a different data subset where sampling is done with replacement. Then, it provides as output the class that is the mode of all tree predictions [77]. In this way, RF overcomes the high sensitivity of decision trees to the training data (i.e., overfitting). The low correlation between models (i.e., randomisation) is the key feature of RF. Trees protect each other from their individual errors as long as they do not constantly make errors in the same direction, and they cooperate to outperform any of the individual constituent models.

AB is an iterative algorithm which uses the boosting method to combine multiple models of the same type, by explicitly seeking models that complement one another [78]. Whereas in bagging approaches individual models are built separately, in boosting each new model is influenced by the performance of those built in the previous iterations. Moreover, AB does not assign an equal weight to all models, but rather creates a set of weighted models, updated at each iteration.

Therefore, the combination of multiple models provides better predictions with respect to simpler learning algorithms, such as SVM, single decision tree (J48), and instance-based learners (k-NN). RF and AB show promising classification performances directly on original unbalanced R_D and E_{NS},

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access



(a) R^* (Original)

(b) R^* (SMOTE)

(c) E_{NS} (Original)

(d) E_{NS} (SMOTE)

FIGURE 4. Classification performance on all datasets

since no noticeable improvements can be appreciated by balancing these datasets using SMOTE.

2) Feature selection

RF and AB algorithms have been tested on R^* and E_{NS} datasets together with different feature selection techniques. In this case, we directly used original unbalanced datasets, since no significant difference in performances has been shown by using SMOTE. Classification results are shown in Figure 5, divided by the reference datasets.

As far as R^* dataset is concerned, RF-GainRatio and RF-PCA provide similar results to those obtained with all the physiological feature set, with information gain ratio as the best metric for feature selection (RF-GainRatio: accuracy= 82.9%, precision= 83.8%, recall = 96.0%, AUPRC = 0.95). Results with AB are similar, showing AB-GainRatio as best performer (AB-GainRatio: accuracy= 85.3%, precision= 85.2%, recall = 97.0%, AUPRC = 0.94). Instead, CFS shows lower performances for both RF and AB. Specifically,it negatively affects more accuracy than precision and recall, thus it reduces specificity.

Classification performances for E_{NS} dataset still indicate information gain ratio as the best feature selection technique for both RF and AB, whereas CFS again shows lower performance than the other two methods (RF-GainRatio: accuracy= 85.3%, precision= 88.5%, recall = 92.0%, AUPRC = 0.98; AB-GainRatio: accuracy= 85.5%, precision= 88.9%, recall = 92.3%, AUPRC = 0.98).

These results demonstrate that it is possible to reach almost equal performances by using a smaller set of input features. This can help discard irrelevant and redundant features, and it can reduce overfitting by providing less opportunity to make predictions based on noisy data.

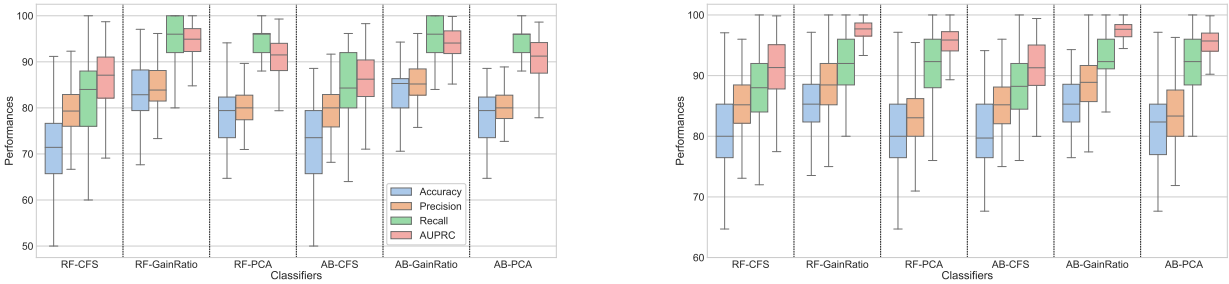
C. TRAINING TIME EVALUATION

Table VIII shows training times (in *s*) for the best-performing configuration of each classifier in R^* and E_{NS}. Training time for k-NN is not reported since it is always less than 0.01 ms, which is much lower than all the other classifiers. This may depend on the fact that it is a heuristic-based classifier that does not require any learning process. RF needs about 1 s for training. Specifically, the measured training time is 1.084 ± 0.159 s for R^* dataset and 0.775 ± 0.136 s for E_{NS}. Instead, the iterative scheme of AB takes a significantly higher training time, which ranges between 26 s and 27 s for both datasets. Finally, BN and J48 take training times less than 0.2 s for both datasets, whereas SVM training time ranges from 1.739 ± 0.434 s for E_{NS} to 11.365 ± 1.026 s for R^*.

Feature selection techniques speed up the training process for the best performing classifiers (i.e., RF and AB), as shown in Table IX. AB training time significantly reduces for both datasets, with AB-GainRatio average training time ranging from 2 s to 4 s. A faster training phase can still be appreciated by applying all the feature selection technique to RF. Specifically, RF-GainRatio training time is 0.797 ± 0.076 s for R^* and 0.523 ± 0.054 s for E_{NS}.

Despite the training time depends on several factors, in-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access



(a) R^* (Original) (b) E_{NS} (Original)

FIGURE 5. RF and AB performance with different feature selection techniques

TABLE VIII. Training time for the full input feature set (seconds)

Dataset	BN	RF	J48	SVM	AB
R^*	0.103 ± 0.075	1.084 ± 0.159	0.169 ± 0.091	11.365 ± 1.026	26.809 ± 0.718
E_{NS}	0.031 ± 0.016	0.775 ± 0.136	0.098 ± 0.029	1.739 ± 0.434	26.495 ± 2.778

TABLE IX. Training time for feature selection (seconds)

Dataset	RF-CFS	RF-GainRatio	RF-PCA	AB-CFS	AB-GainRatio	AB-PCA
R^*	0.786 ± 0.097	0.797 ± 0.076	0.945 ± 0.110	0.211 ± 0.024	2.098 ± 0.750	0.439 ± 0.125
E_{NS}	0.528 ± 0.059	0.523 ± 0.054	0.569 ± 0.074	0.475 ± 0.135	4.020 ± 0.797	5.002 ± 1.088

cluding dataset size, model complexity, algorithm implementation, and also hardware settings, we can state that in our experimental setup best performances can be quickly achieved by using RF, with a training time of approximately 1 s. RF training process becomes even faster using feature selection techniques. Moreover, AB-GainRatio also provides a quite fast training using a smaller set of inputs, while keeping the same classification performances. To speed up the training time represents an even more valuable advantage when training models in mobile settings, using resource-constrained devices.

D. MAIN LIMITATIONS

RF and AB algorithms show promising stress detection results, which support our experimental analysis. This represents a fundamental preliminary assessment of the system's capabilities to correctly classify stressful and non-stressful events in experimental protocols that include different activities. Obtained performances are in line with most binary stress detection studies performed in fully controlled laboratory environments. According to [23], reported accuracy of binary in-lab stress detection studies combining EDA and heart activity ranges from 80.0% to 95.8%, whereas multimodal studies including also different biosignals provide a wider accuracy range ([66%-98%]). Instead, multimodal binary stress detection studies in controlled laboratory environments reviewed in [56] report accuracy values ranging from 79% to 97%.

In order to both train our system with more data and to further test its generalisation capabilities, more subjects and more monitoring sessions are necessary. We are currently working on this in order to use Leave One Subject Out (LOSO) validation scheme to evaluate classification performances on one (or more) new unseen subjects. However, ML algorithms are mainly used for binary or multi-class classification. In order to detect different stress levels, multiple classes should be used, but a higher number of classes can negatively affect the classification performances if not enough patterns for each class are provided. To provide the users, medical specialists, and medical personnel with more useful and detailed information, we need a system able to predict a numerical stress level (SL). This stress score can be computed in an event- or time-based manner (i.e., every minute) to make it timely and easily accessible during the therapy. The inferred SL is a valuable information which can be used to tune and personalise the treatment for frail older subjects, based on their stress-related physiological response. For instance, medical specialists can stop the training if a subject is experiencing a too high stress level for a prolonged time; viceversa, they can push a subject into a more challenging training if the inferred stress level remains low-to-moderate over the time. To this aim, the reference stress ground truth needs to be changed, using stress self-reports and/or clinical assessments at the end of each protocol phase to track individual stress levels throughout the training and rehabilitation activities. Obtained stress scores should be normalized to tackle with the subject-dependent stress perception.

For this reason, in the next section we describe other possible

solutions and we present our proposal for a DSS for online stress monitoring based on a mobile software architecture.

VI. DECISION SUPPORT SYSTEM FOR ONLINE STRESS MONITORING

Despite the wide number of stress detection systems, only few studies present DSS able to support an online stress diagnosis and they are mainly designed for clinical environments. Begum et al. and Nilsson et al. propose two different Case-Based Reasoning (CBR) clinical DSS for stress diagnosis using finger temperature (FT) and Respiratory Sinus Arrhythmia (RSA) patterns, respectively [79], [80]. Features extracted from such patterns are used to formulate a new problem case and, then, they are matched with a case base of collected records using different similarity measures. As any CBR system, they retrieve a ranked list of the most similar records in the case base to propose a suitable solution for the new case. The proposed solutions provide a feedback related to the probability that the subject may suffer from stress-related disorders, but neither of the two approaches aim to infer the user experienced stress level.

The main advantage of CBR systems is that they are very intuitive and do not require any knowledge elicitation to create rules or models, since they reuse the previous knowledge to solve future problems, thus exploiting a reasoning technique very similar to human reasoning [81]. However, CBR systems can require a large storage to build a case base, as well as a large processing time for the similarity matching and case retrieval phase, especially as the case base grows up. Additionally, human interaction is usually necessary to adapt the proposed solution in the reuse and retain phase.

Tartarisco et al. [82] propose an enhanced solution based on a Personal Health System for stress monitoring in natural settings with an additional support to decision making. The proposed sensing system relies on ECG and 3D accelerometer data, but it can be extended to additional wearable and smartphone sensors. The analysis module exploits a supervised SOM for activity recognition, whereas the decision module is a classical FIS, which predicts user stress levels from physiological input features. The proposed FIS structure is defined a priori, thus MFs and rule base are not tuned over input data. This approach does not exploit data modeling and learning techniques, so knowledge elicitation represents a major drawback.

Gaggioli et al. [83] have developed a DSS for real-time stress detection to track users experienced stress during Stress Inoculation Training (SIT) procedures by exploiting the subject's exposure to virtual reality (VR) environments. The DSS integrates ECG, EEG, breathing rate (BR), and body gestures collected using multiple sensing platforms in order to extract several features, and it is built by using fuzzy logic in conjunction with supervised SOMs. In the training phase, SOMs are trained using fuzzified physiological features along with self- and clinical-reported stress levels, which are used as ground truth during VR exposure. In the test phase, the trained DSS predicts a stress level whenever a new instance is provided. This solution relies on a supervised variant of SOMs, which still needs self-reports and/or clinical annotations as stress ground truth, whereas membership functions for each feature are defined a priori. Moreover, SOMs require further analysis of the output patterns, but no precise indication about how the stress score is computed from SOMs is provided.

To overcome these limitations, we propose a mobile architecture for a novel DSS aimed at online stress monitoring and stress level detection. The system architecture is presented in Figure 6. It consists of different wireless modules, integrated to perform data collection, processing, analysis and support to decision making. Specifically, the system is equipped with:

1) a sensing platform, composed by different heterogeneous wearable devices;
2) a mobile node, which collects and processes physiological data from the sensing sources;
3) a back-end node, which hosts the data analysis and decision module.

The m-health solution used in the pilot study and presented in section III-B will be further enhanced with data preprocessing and feature extraction functionalities, following the principles of edge computing [84]. Moreover, we are currently implementing the decision module by investigating different predictive models that may provide an accurate stress inference. We plan to maintain storage and computational intensive features on the back-end module as a preliminary solution, thus respecting the real-time constraints of the DSS. In a second phase, we will also study the feasibility of porting the trained model on the mobile node, and also to train and update the model directly in mobile settings.

The system have also a closed-loop structure, by integrating self-reports and/or clinical evaluations collected at the end of each protocol phase to be used as ground truth to track user perceived stress levels throughout the training sessions.

In the case of MCI frail older people, it is reasonable that clinical evaluations are used to track the user stress level during the therapy. In addition, medical specialists generally supervise the rehabilitation activities, and to provide their clinical feedback can be considered as a natural step to verify and support the DSS.

In terms of physiological data analysis, features are extracted on a time-based manner by choosing 1-minute overlapped time windows as decision time (T_D), but this is a configurable choice aimed at providing a close stress tracking during the training, while also allowing to extract meaningful features within a short time period, especially for stress-related HRV features.

Given the versatility of the proposed architecture, this system can accommodate several predictive models, able to receive as input the physiological features computed on the mobile node and to provide as output an inferred SL. Multiple models could also coexist and cooperate as an ensemble to improve predictive performances. As shown in section V, RF and AB turned out to be the best performing learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2985301, IEEE Access

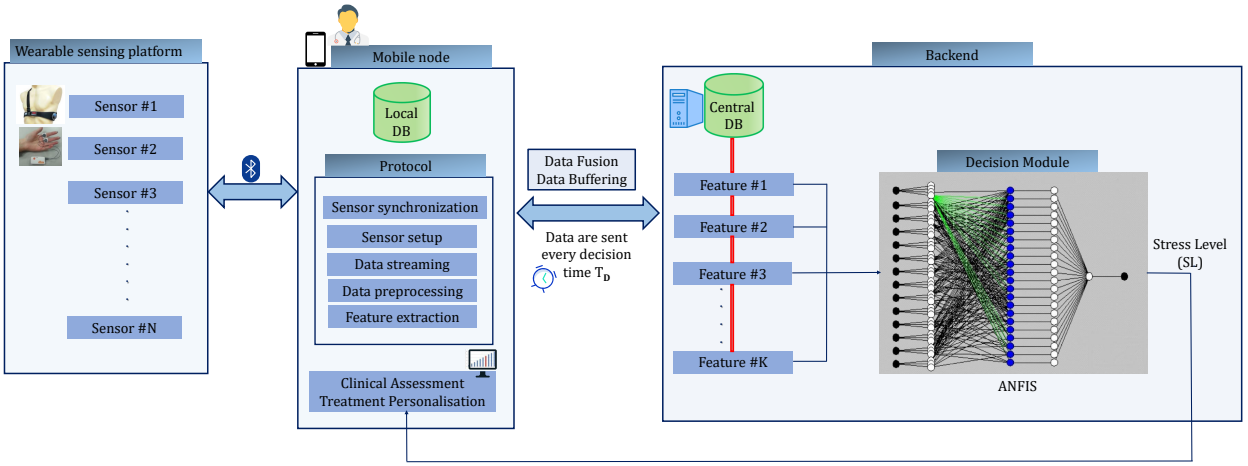


FIGURE 6. Decision Support System for online stress monitoring based on a mobile architecture

scheme for binary stress classification in our study. Thus, bagged and boosted regression tree ensemble, may still be used for predicting perceived stress scores. ANNs, such as Long Short-Term Memory networks (LSTM), represent another solution to work with sequence and time series data for regression tasks. However, loss of interpretability is a drawback when applying ANNs or ensemble learning and this can make knowledge representation very difficult to obtain an intelligible model structure in order to understand the rationale behind a given prediction.

Adaptive Neuro Fuzzy Inference Systems (ANFISs) may simplify the process of knowledge representation through the intrinsic explanatory nature of fuzzy rules [85], [86]. ANFIS provides a method for the fuzzy modeling procedure which shares learning and computational capabilities typical of deep ANNs to build a FIS, whose structure and parameters (MFs and rule base) are automatically learned and tuned to create a model upon input/output data. Instead of defining FIS structure based on a trial and error strategy or gaining a knowledge base from domain experts, ANFIS computes MF parameters and define rules that best allow the associated fuzzy inference engine to learn from the data it is modeling. Parameter tuning can be performed by applying back propagation gradient descent (BPGD) method or other hybrid optimizations. By exploiting MFs and fuzzy rules, ANFIS maintains explanatory power and is able to reveal the functionalities stored into the learned model and used in the decision process, especially if the system has been designed with a limited number of input variables. Feature selection and reduction techniques reported in section IV, as well as genetic algorithms and other heuristic search approaches, may be exploited for input selection. Moreover, a reduced input set can be a key point when extracting input features and training models in mobile settings.

Since all the aforementioned approaches are data-driven techniques for modeling or model-following, a system cold start is needed to train each model. The training set can be built by collecting physiological data, along with clinical-reported stress scores, for a first training session for each of the involved subjects. This system can be used to define a user personalised model by using only individual data, whereas a generalized model can be built collecting data from different subjects.

Finally, to evaluate the generalization capabilities of the proposed system, we intend to apply it to new unseen training sessions by extending the pilot study and including additional monitoring sessions for each subject by modifying the rehabilitation sessions, both in terms of cognitive and physical activities. This will allow us to further investigate the effect of different training procedure on MCI frail older subjects and provide a performance analysis of the proposed DSS for online stress monitoring.

VII. CONCLUSIONS

Cognitive training, as well as other rehabilitation activities, are fundamental for MCI frail older adults to maintain their abilities and a good quality of life. However, this subject category is also usually affected by a chronic stress condition, which can be further worsened by the requested activity. Our study focuses on the analysis of the impact of a specific training protocol on the cognitive performances and the stress response of a group of MCI frail older adults through a m-health solution and wearable physiological sensors. The results of the pilot study show an evident improvement of the cognitive outcomes after the proposed physical and cognitive training, and the subjects present an adaptive stress response to the requested activity. In addition, we used the collected dataset to evaluate the performances of a stress detection system based on 6 different ML algorithms, showing promising results with respect to the performances obtained in laboratory settings with predefined stimuli. This analysis introduces our proposal of a novel technological solution aimed

at supporting the medical specialist to define personalised training sessions based on online stress monitoring. Since the subjects and the LTC facility demonstrated a great interest in the proposed solution and they agreed to participate to new monitoring sessions on a periodic time basis, we are integrating the new DSS system for online stress monitoring in the m-health solution in order to: (i) study the impact of the training sessions on an individual basis, (ii) validate the proposed solutions in terms of accuracy of the detected stress level, and (iii) effectively analyse the personalisation benefits on the final user on a long-term basis. The active collaboration with the LTC facility will also allow us to enlarge the study and further support the presented results.

ACKNOWLEDGMENTS

The authors wish to thank iCARE s.r.l., all the nursing home personnel and guests for their collaboration in the experimental phase.

REFERENCES

[1] Q.-L. Xue, "The frailty syndrome: definition and natural history," Clinics in geriatric medicine, vol. 27, no. 1, pp. 1–15, 2011.

[2] F. Blaskovits, J. Tyerman, and M. Luctkar-Flude, "Effectiveness of neurofeedback therapy for anxiety and stress in adults living with a chronic illness: a systematic review protocol," JBI database of systematic reviews and implementation reports, vol. 15, no. 7, pp. 1765–1769, 2017.

[3] C. U. Onyike, J.-M. E. Sheppard, J. T. Tschanz, M. C. Norton, R. C. Green, M. Steinberg, K. A. Welsh-Bohmer, J. C. Breitner, and C. G. Lyketsos, "Epidemiology of apathy in older adults: the cache county study," The American Journal of Geriatric Psychiatry, vol. 15, no. 5, pp. 365–375, 2007.

[4] T. Sugimoto, T. Sakurai, R. Ono, A. Kimura, N. Saji, S. Niida, K. Toba, L.-K. Chen, and H. Arai, "Epidemiological and clinical significance of cognitive frailty: a mini review," Ageing research reviews, vol. 44, pp. 1–7, 2018.

[5] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, "Mild cognitive impairment: a concept in evolution," Journal of internal medicine, no. 3, pp. 214–228, 2014.

[6] Q. Hao, B. Dong, M. Yang, B. Dong, and Y. Wei, "Frailty and cognitive impairment in predicting mortality among oldest-old people," Frontiers in aging neuroscience, vol. 10, 2018.

[7] A. Diamond, "Executive functions," Annual review of psychology, vol. 64, pp. 135–168, 2013.

[8] M. E. Kelly, D. Loughrey, B. A. Lawlor, I. H. Robertson, C. Walsh, and S. Brennan, "The impact of exercise on the cognitive functioning of healthy older adults: a systematic review and meta-analysis," Ageing research reviews, vol. 16, pp. 12–31, 2014.

[9] C. Hammen, E. Y. Kim, N. K. Eberhart, and P. A. Brennan, "Chronic and acute stress and the prediction of major depression in women," Depression and anxiety, vol. 26, no. 8, pp. 718–723, 2009.

[10] S. Cohen, D. Janicki-Deverts, and G. E. Miller, "Psychological stress and disease," Jama, vol. 298, no. 14, pp. 1685–1687, 2007.

[11] G. G. Ortiz, F. P. P. Moisés, M. Mireles-Ramírez, L. J. Flores-Alvarado, H. González-Usigli, V. J. Sanchez-Gonzalez, A. L. Sanchez-Lopez, L. Sánchez-Romero, E. I. Díaz-Barba, J. F. Santoscoy-Gutiérrez et al., "Oxidative stress: Love and hate history in central nervous system," in Advances in protein chemistry and structural biology. Elsevier, 2017, vol. 108, pp. 1–31.

[12] W. R. Lovallo, Stress and health: Biological and psychological interactions. Sage publications, 2015.

[13] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," Computer methods and programs in biomedicine, vol. 108, no. 3, pp. 1287–1301, 2012.

[14] D. Mannarelli, C. Pauletti, P. Mancini, A. Fioretti, A. Greco, M. De Vincentiis, and F. Fattapposta, "Selective attentional impairment in chronic tinnitus: Evidence from an event-related potentials study," Clinical Neurophysiology, vol. 128, no. 3, pp. 411–417, 2017.

[15] M. Angevaren, G. Aufdemkampe, H. Verhaar, A. Aleman, L. Vanhees et al., "Physical activity and enhanced fitness to improve cognitive function in older people without known cognitive impairment," Cochrane Database Syst Rev, vol. 3, no. 3, pp. 1–73, 2008.

[16] P. J. Smith, J. A. Blumenthal, B. M. Hoffman, H. Cooper, T. A. Strauman, K. Welsh-Bohmer, J. N. Browndyke, and A. Sherwood, "Aerobic exercise and neurocognitive performance: a meta-analytic review of randomized controlled trials," Psychosomatic medicine, vol. 72, no. 3, p. 239, 2010.

[17] D. Song, S. Doris, P. W. Li, and Y. Lei, "The effectiveness of physical exercise on cognitive and psychological outcomes in individuals with mild cognitive impairment: A systematic review and meta-analysis," International journal of nursing studies, 2018.

[18] J. G. Van Uffelen, M. J. Chinapaw, W. van Mechelen, and M. Hopman-Rock, "Walking or vitamin b for cognition in older adults with mild cognitive impairment? a randomized controlled trial," British journal of sports medicine, 2008.

[19] L. D. Baker, L. L. Frank, K. Foster-Schubert, P. S. Green, C. W. Wilkinson, A. McTiernan, S. R. Plymate, M. A. Fishel, G. S. Watson, B. A. Cholerton et al., "Effects of aerobic exercise on mild cognitive impairment: a controlled trial," Archives of neurology, vol. 67, no. 1, pp. 71–79, 2010.

[20] L. S. Nagamatsu, A. Chan, J. C. Davis, B. L. Beattie, P. Graf, M. W. Voss, D. Sharma, and T. Liu-Ambrose, "Physical activity improves verbal and spatial memory in older adults with probable mild cognitive impairment: a 6-month randomized controlled trial," Journal of aging research, vol. 2013, 2013.

[21] T. Suzuki, H. Shimada, H. Makizako, T. Doi, D. Yoshida, K. Ito, H. Shimokata, Y. Washimi, H. Endo, and T. Kato, "A randomized controlled trial of multicomponent exercise in older adults with mild cognitive impairment," PloS one, vol. 8, no. 4, p. e61483, 2013.

[22] L. C.-w. Lam, W. C. Chan, T. Leung, A. W.-t. Fung, and E. M.-f. Leung, "Would older adults with mild cognitive impairment adhere to and benefit from a structured lifestyle activity intervention to enhance cognition?: a cluster randomized controlled trial," PLoS One, vol. 10, no. 3, p. e0118173, 2015.

[23] 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, p. 103139, 2019.

[24] T. Reinhardt, C. Schmahl, S. Wüst, and M. Bohus, "Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the mannheim multicomponent stress test (mmst)," Psychiatry research, vol. 198, no. 1, pp. 106–111, 2012.

[25] S. Betti, R. M. Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. Del Ry, and F. Cavallo, "Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers," IEEE Transactions on Biomedical Engineering, vol. 65, no. 8, pp. 1748–1758, 2018.

[26] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in International Conference on Mobile Computing, Applications, and Services. Springer, 2010, pp. 282–301.

[27] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva, "Using activity-related behavioural features towards more effective automatic stress detection," PloS one, vol. 7, no. 9, p. e43571, 2012.

[28] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis, "Stress and anxiety detection using facial cues from videos," Biomedical Signal Processing and Control, vol. 31, pp. 89–101, 2017.

[29] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on. IEEE, 2013, pp. 209–214.

[30] L. Torres-Salomao, M. Mahfouf, and E. El-Samahy, "Pupil diameter size marker for incremental mental stress detection," in 2015 17th international conference on e-health networking, application & services (HealthCom). IEEE, 2015, pp. 286–291.

[31] M. X. Huang, J. Li, G. Ngai, and H. V. Leong, "Stressclick: Sensing stress from gaze-click patterns," in Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 1395–1404.

[32] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier, "Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings," in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2017, pp. 456–463.

[33] M. N. H. Mohd, M. Kashima, K. Sato, and M. Watanabe, "Mental stress recognition based on non-invasive and non-contact measurement from stereo thermal and visible sensors," International Journal of Affective Engineering, vol. 14, no. 1, pp. 9–17, 2015.

[34] X. Zhang, Y. Lyu, X. Luo, J. Zhang, C. Yu, H. Yin, and Y. Shi, "Touch sense: Touch screen based mental stress sense," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 2, pp. 1–18, 2018.

[35] Y. Cho, S. J. Julier, and N. Bianchi-Berthouze, "Instant stress: Detection of perceived mental stress through smartphone photoplethysmography and thermal imaging," JMIR mental health, vol. 6, no. 4, p. e10140, 2019.

[36] T. Chen, P. Yuen, M. Richardson, G. Liu, and Z. She, "Detection of psychological stress using a hyperspectral imaging technique," IEEE Transactions on Affective Computing, vol. 5, no. 4, pp. 391–405, 2014.

[37] S. D. Gunawardhane, P. M. De Silva, D. S. Kulathunga, and S. M. Arunatileka, "Non invasive human stress detection using key stroke dynamics and pattern variations," in 2013 International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2013, pp. 240–247.

[38] E. A. Sağbaş, S. Korukoglu, and S. Balli, "Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques," Journal of Medical Systems, vol. 44, no. 4, pp. 1–12, 2020.

[39] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," Neuropsychobiology, vol. 28, no. 1-2, pp. 76–81, 1993.

[40] F. Scarpina and S. Tagini, "The stroop color and word test," 2017.

[41] A.-M. Brouwer and M. A. Hogervorst, "A new paradigm to induce mental stress: the sing-a-song stress test (ssst)," Frontiers in neuroscience, vol. 8, p. 224, 2014.

[42] D. Lowery, R. B. Fillingim, and R. A. Wright, "Sex differences and incentive effects on perceptual and cardiovascular responses to cold pressor pain," Psychosomatic medicine, vol. 65, no. 2, pp. 284–291, 2003.

[43] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic stress detection in working environments from smartphones' accelerometer data: a first step," IEEE journal of biomedical and health informatics, vol. 20, no. 4, pp. 1053–1060, 2015.

[44] A. Muaremi, B. Arnrich, and G. Tröster, "Towards measuring stress with smartphones and wearable devices during workday and sleep," BioNanoScience, vol. 3, no. 2, pp. 172–183, 2013.

[45] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," Sensors, vol. 19, no. 8, p. 1849, 2019.

[46] B. Egilmez, E. Poyraz, W. Zhou, G. Memik, P. Dinda, and N. Alshurafa, "Ustress: Understanding college student subjective stress using wrist-based passive sensing," in 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, 2017, pp. 673–678.

[47] L.-l. Chen, Y. Zhao, P.-f. Ye, J. Zhang, and J.-z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," Expert Systems with Applications, vol. 85, pp. 279–291, 2017.

[48] P. Zontone, A. Affanni, R. Bernardini, A. Piras, and R. Rinaldo, "Stress detection through electrodermal activity (eda) and electrocardiogram (ecg) analysis in car drivers," in 2019 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019, pp. 1–5.

[49] E. Vildjiounaite, J. Kallio, V. Kyllönen, M. Nieminen, I. Määttänen, M. Lindholm, J. Mäntyjärvi, and G. Gimel'farb, "Unobtrusive stress detection on the basis of smartphone usage data," Personal and Ubiquitous Computing, vol. 22, no. 4, pp. 671–688, 2018.

[50] Z. D. King, J. Moskowitz, B. Egilmez, S. Zhang, L. Zhang, M. Bass, J. Rogers, R. Ghaffari, L. Wakschlag, and N. Alshurafa, "micro-stress ema: A passive sensing framework for detecting in-the-wild stress in pregnant mothers," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, no. 3, p. 91, 2019.

[51] S. Cohen, T. Kamarck, R. Mermelstein et al., "Perceived stress scale," Measuring stress: A guide for health and social scientists, vol. 10, 1994.

[52] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," Annu. Rev. Clin. Psychol., vol. 4, pp. 1–32, 2008.

[53] S. I. Donaldson and E. J. Grant-Vallone, "Understanding self-report bias in organizational behavior research," Journal of business and Psychology, vol. 17, no. 2, pp. 245–260, 2002.

[54] B. Kikhia, T. G. Stavropoulos, S. Andreadis, N. Karvonen, I. Kompatsiaris, S. Sävenstedt, M. Pijl, and C. Melander, "Utilizing a wristband sensor to measure the stress level for people with dementia," Sensors, vol. 16, no. 12, p. 1989, 2016.

[55] D. L. Algase, E. R. Beattie, S. A. Leitsch, and C. A. Beel-Bates, "Biomechanical activity devices to index wandering behaviour in dementia," American Journal of Alzheimer's Disease & Other Dementias®, vol. 18, no. 2, pp. 85–92, 2003.

[56] E. Smets, W. De Raedt, and C. Van Hoof, "Into the wild: The challenges of physiological stress detection in laboratory and ambulatory settings," IEEE journal of biomedical and health informatics, vol. 23, no. 2, pp. 463–473, 2018.

[57] A. de Santos Sierra, C. S. Ávila, J. G. Casanova, and G. B. del Pozo, "A stress-detection system based on physiological signals and fuzzy logic," IEEE Transactions on Industrial Electronics, vol. 58, no. 10, pp. 4857–4865, 2011.

[58] A. Salazar-Ramirez, E. Irigoyen, R. Martinez, and U. Zalabarria, "An enhanced fuzzy algorithm based on advanced signal processing for identification of stress," Neurocomputing, vol. 271, pp. 48–57, 2018.

[59] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-based analysis for personalized stress evaluation using physiological signals," IEEE journal of biomedical and health informatics, vol. 19, no. 1, pp. 275–281, 2014.

[60] B. Arnrich, C. Setz, R. La Marca, G. Tröster, and U. Ehlert, "What does your chair know about your stress level?" IEEE Trans. Information Technology in Biomedicine, vol. 14, no. 2, pp. 207–214, 2010.

[61] D. Huysmans, E. Smets, W. De Raedt, C. Van Hoof, K. Bogaerts, I. Van Diest, and D. Helic, "Unsupervised learning for mental stress detection," in Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies, vol. 4, 2018, pp. 26–35.

[62] A. Sydor, "Conducting research into hidden or hard-to-reach populations," Nurse researcher, vol. 20, no. 3, 2013.

[63] F. Delmastro, C. Dolciotti, F. Palumbo, M. Magrini, F. D. Martino, D. L. Rosa, and U. Barcaro, "Long-term care: how to improve the quality of life with mobile and e-health services," in 14th International Conference on Wireless and Mobile Computing, Networking and Communications, WiMob 2018, Limassol, Cyprus, October 15-17, 2018, 2018, pp. 12–19.

[64] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. Kleiger et al., "Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology," 1996.

[65] W. Bouaziz, E. Schmitt, G. Kaltenbach, B. Geny, and T. Vogel, "Health benefits of cycle ergometer training for older adults over 70: a review," European Review of Aging and Physical Activity, vol. 12, no. 1, p. 8, 2015.

[66] P. Caffarra, G. Vezzadini, F. Dieci, F. Zonato, and A. Venneri, "A short version of the stroop test: normative data in an italian population sample," Nuova Rivista di Neurologia, vol. 12, no. 4, pp. 111–115, 2002.

[67] J. H. Horne and S. L. Baliunas, "A prescription for period analysis of unevenly sampled time series," The Astrophysical Journal, vol. 302, pp. 757–763, 1986.

[68] S. for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson, and D. L. Filion, "Publication recommendations for electrodermal measurements," Psychophysiology, vol. 49, no. 8, pp. 1017–1034, 2012.

[69] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," Journal of neuroscience methods, vol. 190, no. 1, pp. 80–91, 2010.

[70] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," in 2007 29th annual international conference of the ieee engineering in medicine and biology society. IEEE, 2007, pp. 4656–4659.

[71] E. Smets, P. Casale, U. Großekathöfer, B. Lamichhane, W. De Raedt, K. Bogaerts, I. Van Diest, and C. Van Hoof, "Comparison of machine learning techniques for psychophysiological stress detection," in International Symposium on Pervasive Computing Paradigms for Mental Health. Springer, 2015, pp. 13–22.

[72] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," PloS one, vol. 10, no. 3, p. e0118432, 2015.

[73] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.

[74] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[75] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," International Journal of Information Technology and Knowledge Management, vol. 2, no. 2, pp. 271–277, 2010.
[76] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," in 2010 international conference on system science, engineering design and manufacturing informatization, vol. 1. IEEE, 2010, pp. 27–30.
[77] I. Barandiaran, "The random subspace method for constructing decision forests," IEEE Trans. Pattern Anal. Mach. Intell, vol. 20, no. 8, pp. 1–22, 1998.
[78] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[79] S. Begum, M. U. Ahmed, P. Funk, N. Xiong, and B. Von Schéele, "A case-based decision support system for individual stress diagnosis using fuzzy similarity matching," Computational Intelligence, vol. 25, no. 3, pp. 180–195, 2009.
[80] M. Nilsson, P. Funk, E. M. Olsson, B. von Schéele, and N. Xiong, "Clinical decision-support for diagnosing stress-related disorders by applying psychophysiological medical knowledge to an instance-based learning system," Artificial Intelligence in Medicine, vol. 36, no. 2, pp. 159–176, 2006.
[81] N. Shoaip, S. El-Sappagh, S. Barakat, and M. Elmogy, "Reasoning methodologies in clinical decision support systems: A literature review," in U-Healthcare Monitoring Systems. Elsevier, 2019, pp. 61–87.
[82] G. Tartarisco, G. Baldus, D. Corda, R. Raso, A. Arnao, M. Ferro, A. Gaggioli, and G. Pioggia, "Personal health system architecture for stress monitoring and support to clinical decisions," Computer Communications, vol. 35, no. 11, pp. 1296–1305, 2012.
[83] A. Gaggioli, P. Cipresso, S. Serino, G. Pioggia, G. Tartarisco, G. Baldus, D. Corda, M. Ferro, N. Carbonaro, A. Tognetti et al., "A decision support system for real-time stress detection during virtual reality exposure." in MMVR, 2014, pp. 114–120.
[84] R. Buyya and S. N. Srirama, Exploiting Fog Computing in Health Monitoring. Wiley, 2019, pp. 291–318. [Online]. Available: https://ieeexplore.ieee.org/document/8654143
[85] J.-S. R. Jang et al., "Fuzzy modeling using generalized neural networks and kalman filter algorithm." in AAAI, vol. 91, 1991, pp. 762–767.
[86] J.-S. Jang, "Anfis: adaptive-network-based fuzzy inference system," IEEE transactions on systems, man, and cybernetics, vol. 23, no. 3, pp. 665–685, 1993.

FRANCA DELMASTRO is a researcher of the IIT Institute of the National Research Council of Italy (CNR). Her research interests are in the field of pervasive and ubiquitous computing with specific attention to opportunistic and wearable sensing, personalised m-health solutions, context- and social-aware middleware and mobile applications and recommender systems. She published several works on international journals, conferences and book chapters in the pervasive and mobile computing area. She participated in the organization of several international conferences. She served as TPC Vice Chair of IEEE PerCom 2017, Workshops co-Chair of IEEE PerCom 2014, General Co-Chair of Work-in-Progress session of IEEE PerCom 2012, and General Chair of several international workshops. She has been Guest Editor of Special Issues on Pervasive Healthcare for Elsevier Pervasive and Mobile Computing Journal and Computer Communication Journal. She is a National Expert for ICT solutions for health and well-being in the framework of EU H2020 program.

FLAVIO DI MARTINO is a Ph.D. student in Computer Engineering at the University of Pisa in collaboration with the IIT Institute of the National Research Council of Italy (CNR). He received a M.S. degree in Biomedical Engineering from the University of Pisa in 2015. His research activity is related to the study and application of m-health solutions for physiological and behavioural monitoring, predictive models, and decision support systems for human behavioural analytics.

CRISTINA DOLCIOTTI is sub-investigator and data manager in clinical trials at Department of Neuroscience of the University of Pisa (Italy), and consultant physiatrist and nutritionist at iCARE Heath Facility in Viareggio (Italy). She received a M.S. degree in Medicine and Surgery from the University of Pisa in 2006. She is specialized in Dietetics, Neurorehabilitation and Clinical Posturology. She worked at complex Unit of Neurology of Versilia Hospital (Viareggio, Italy) from 2006 to 2011, at Neurorehabilitation Unit of Cisanello Hospital (Pisa, Italy) from 2015 to 2016, and at Complex Unit of Neurology at New Hospital of Apuane (Massa-Carrara, Italy) from 2016 to 2018. Her research activity is related to neurodegenerative diseases, neuropsychology and pathophysiology of aging. She is author of 20 peer-reviewed publications.