Challenges of learning human digital twin: case study of mental wellbeing

Using sensor data and machine learning to create HDT

Elena Vildjiounaite
VTT Technical Research Centre of Finland
elena.vildjiounaite@vtt.fi

Johanna Kallio
VTT Technical Research Centre of Finland
johanna.kallio@vtt.fi

Julia Kantorovitch
VTT Technical Research Centre of Finland
julia.kantorovitch@vtt.fi

Atte Kinnula
VTT Technical Research Centre of Finland
atte.kinnula@vtt.fi

Simão Ferreira
Center for Translational Health and Medical Biotechnology Research
School of Health of Polytechnic Institute of Porto
sprf@ess.ipp.pt

Matilde A. Rodrigues
Center for Translational Health and Medical Biotechnology Research
School of Health of Polytechnic Institute of Porto
mar@ess.ipp.pt

Nuno Rocha
Center for Translational Health and Medical Biotechnology Research
School of Health of Polytechnic Institute of Porto
nrocha@ess.ipp.pt

ABSTRACT

Human Digital Twin (HDT) is a powerful tool to create a virtual replica of a human, to be used for example for designing interactions with physical systems, preventing cognitive overload, managing human capital, and maintaining a healthy and motivated workforce. Building human twins is a challenging task due to the need to reliably represent each corresponding human being, and the fact that human beings notably differ from each other. Therefore, relying solely on expert knowledge is insufficient, and human twins must learn the specifics of each individual in order to accurately represent them. This paper focuses on AI methods for modelling the mental wellbeing of knowledge workers because the mounting cognitive demands of both white-collar and blue-collar work lead to employees' stress, and stress leads to diminished creativity and motivation, increased sick leaves, and in severe cases, accidents, burnouts, and disabilities. This paper describes the main building blocks of AI-based detectors of mental stress and highlights the main challenges and future directions of research., which are expected to be relevant also for HDT learning in other domains because the high degree of individuality is ubiquitous in all human activities.

PETRA '23, July 05–07, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07.
https://doi.org/10.1145/3594806.3596538

CCS CONCEPTS

• **Ubiquitous and mobile computing**; • **Artificial Intelligence**; • **User characteristics**;

KEYWORDS

Personalisation, Mental Wellbeing, Stress, Workplace, Human Digital Twin

ACM Reference Format:
Elena Vildjiounaite, Johanna Kallio, Julia Kantorovitch, Atte Kinnula, Simão Ferreira, Matilde A. Rodrigues, and Nuno Rocha. 2023. Challenges of learning human digital twin: case study of mental wellbeing: Using sensor data and machine learning to create HDT. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594806.3596538

1 INTRODUCTION

AI-based human modelling is an active research area, with studies aiming at learning preferences of humans (e.g., to use in recommender systems and for user interface adaptation) and human activities (e.g., to assist workers and dementia sufferers when they do not know what to do next). Many studies aim also at recognising human cognitive and mental problems, both for medical reasons and for ordinary life. For example, detection of students' stress during online lessons may help to adapt study programme; detection of employees' stress may help to improve work culture, employee motivation and health, whereas workplace problems can cause accidents, employee sicknesses, turnover, and early retirement. To ensure successful collaboration between humans and AI in so-called

hybrid intelligence systems, it is also important to recognise human stress, shortcomings and preferences, and to adapt interaction accordingly.

A big problem in all these domains is human diversity, which makes "one model suits all" approach infeasible. Factors, that threaten mental wellbeing of each person (i.e., reasons for his/ her cognitive and physical overload, fatigue, boredom etc.), notably differ between different individuals. For example, tasks, which some humans perceive as easy, may cause stress in other humans; tasks, which some humans perceive as interesting, may be perceived by others as boring - and again, stressful, because insufficient variety of work can also cause stress [1]. (In this work we will use the term "stress" in a broad sense, such as "a state of mental tension and worry caused by problems in your life, work, etc."[1]).

Digital Twins are a useful concept for this, as they have a 1-to-1 relationship with an instance of a specific real-world entity [2]. Obtaining digital twins of each employee solely via questionnaires would be infeasible, however, because human beings may lack understanding of reasons for their stress, or may not be honest, for example, because of fear of unemployment. Thus intelligent systems should be able to learn by observing humans and their context. AI methods to detect stress and stressors (i.e., factors that may cause stress) from sensor data is an active research area now, but this area has many research challenges. This work outlines the challenges and future research directions on the example of stress detection studies with students and knowledge workers, but they can be generalised to other domains: for example, manufacturing industries present increasingly more demands for cognitive abilities of the employees. Currently, however, stress detection in manufacturing industries is less advanced than in the domain of knowledge work [3].

First challenge in building human digital twin is to obtain observations and the corresponding ground truth. Observations are sensor data, such as data from wearable devices, mobile phones, computers, video cameras etc., and observations should be continuous because digital twins must evolve together with the humans they represent [4]. Ground truth in stress studies can be notes of external observers, subjects' biomarkers (e.g., cortisol) and self-reports of the study subjects. In long-term real-life data collections usually only self-reports of the subjects can be obtained, and even this is not so easy because of reporting fatigue. Thus it is necessary to develop AI methods to learn human models and to update them when humans evolve, not requiring large quantities of self-reports, and to develop convenient and acceptable for humans sensor-based monitoring systems and applications to obtain self-reports.

To date, research into these problems is only emerging. The overwhelming majority of stress detection studies are conducted in the laboratories, where studies are usually short (a few hours) and therefore focus on detection of acute stress. For example, typical lab conditions to induce stress are requirements to perform mental arithmetic quickly and/ or to write a text despite interruptions. In real life, however, knowledge workers rarely perform mental arithmetic and can disable email notifications to focus on writing. Hence it is not clear to which extent lab studies can help to prevent employee turnover or burnout in real work, moreover that in real work stress can be due to social relations, poor work-life balance, mismatch between personal expectations and results, and even boredom, depending on personality [1][5]. On the other hand, time pressure, interruptions and multitasking can also play positive role. For example, during 10 months long study, described in [6], the test subjects were asked to provide on a daily basis self-reports and, whenever possible, also free-form comments. Not stressful days were often accompanied by the comments that subjects had to perform difficult tasks on these days, but were "inspired and did a lot", whereas some stressful days were accompanied by an explanation "I tried to write, but I wasn't inspired".

Since real life is very different from lab studies, we believe that it is necessary to learn human digital twins from real life data. This paper is organised as follows: next section presents state of the art in real life stress detection studies, Section 3 presents building blocks of AI systems to learn human models, associated challenges and future directions. Section 4 presents summary and conclusion.

2 STATE OF THE ART IN AI-BASED HUMAN STRESS MODELLING

2.1 Database searches

Since we performed extensive literature searches for our previous publications, such as [6][7], for this paper we only searched for the recent studies. First, we searched ScienceDirect. In order not to miss any recent developments, we conducted a very broad search: by typing "stress detection" in the search window and by filtering the results to include only recent works (years 2022, 2023, 2024). Based on previous experiences, we did not use other filters in ScienceDirect because in the past, filters removed relevant papers. This search returned over 10000 results, and we checked the first 500 titles. Among them we found 16 mental condition assessment studies, relevant to our goal of detecting stress of ordinary employees (not patients, not pilots etc.) for the goal to improve their mental wellbeing, nine of which used real life data. Six of these works, however, were analysing text, which creates high privacy threats, and one study recorded data of students for a very short time: two hours before the exam and then for a short time after holidays. The latter approach seems to be very common, because a recent review of stress detection in students reviewed 44 articles and concluded that although stressors in many studies were real, measurements were usually performed in the labs and thus the main challenge is "to bring detection systems into the classroom" [8]. One more study aimed at detecting stressful situations in manufacturing: the system was first trained and tested in the lab with five subjects and then tested in a field during one shift. System accuracy in a lab was reported as 76%, whereas about real work performance the authors concluded that "from a preliminary assessment, it appears that the software framework can cluster rest conditions and potential stress conditions for the user." Whether potential stress conditions indeed caused stress, was not reported. In our view, this confirms difficulties of developing and testing stress detectors in real life and importance of finding ways to do it. Hence only one of the found studies contributed to our summary Table 1.

Next, we conducted search in Springer database by typing the same expression "stress detection" in the search window. We filtered the results by years (2022 and beyond). Then we filtered results

[1]https://www.britannica.com/dictionary/stress

to belong to the area of Computer Science and to include only journal articles. We got about 1000 journal papers, checked first 200 titles and found only one paper which employed real life data in the form of facial images, obtained from the web, but these data were used only for system training, whereas testing was done only using acted video recording of various emotional expressions [9]. A similar search in IEEE Explore returned 278 journal articles; we browsed first 200 titles and found one study that analysed sleep; one study that analysed YouTube videos to recognise arousal/ valence, but not stress; and one field study. Below we describe the found long-term real life studies.

2.2 Summary of field studies

First stress detection studies, performed in the labs, demonstrated that stress can be detected from physiological data, for example, heart rate variability (HRV). Hence first field studies attempted at transferring to real life methods, developed in laboratory studies, but very soon found that physiological data are notably affected by daily life activities, such as motion, eating, drinking, caffeine intake, conversation and motion [10] and that test subjects are not always happy to wear sensors. For example, Muaremi et al. [11] attempted at collecting HRV data during bedtimes in addition to mobile phone data, collected during daytimes, but even in this setup 12 subjects did not wear physiological sensors for more than one night.

Many other real life studies employed mobile phone as the only sensor [12][13][14][15][16][17][18][19] or in combination with wearable devices [20][21]. These studies employed different types of information from the phones: mobile phone usage data (e.g., applications and interaction types, such tap, scroll, swipe, and text writing), physical activity data (e.g., phone acceleration, user locations) and social activity. These data are called behavioural data. Phone condition also appeared to be useful for stress detection, e.g., temperature of batteries, number of operating apps and frequency of changing display [16].

Use of mobile phones in stress detection does not bother end users with the need to wear and charge any extra gadgets, but data collection drains phone battery. Use of environmental sensors instead of, or in addition to, mobile phones would help to reduce phone battery consumption and/ or to increase stress detection accuracy. Recent real-life studies, performed to date with environmental sensors, include analysis of motion trajectories, obtained from depth cameras [6] and analysis of indoor environmental quality data [7]. In addition, there were numerous short-term studies into stress detection via analysis of computer data (keyboard/ mouse), for example, lab study in [22] and real life study into stress of students during exams [23], and lab studies with various kinds of other sensors, e.g., depth and thermal cameras and pressure sensors.

Stress detection models can be (1) general, i.e., trained on data of multiple subjects; (2) cluster-based, i.e., trained on data of similar subjects (which requires a good metrics of similarity between humans and (3) person-specific, i.e., trained using only data of each end user. Due to differences between human beings, person-specific models are usually the most accurate, provided that the end users supply large numbers of stress labels to train the system (stress label is information whether a person was feeling stressed or not and (possibly) information regarding stress severity). Average differences between accuracies of person-specific and general models exceeded 20% in cases of using mobile phone usage data in [13], accelerometer data in [14], posture data in [24] and keyboard/ mouse data in [23], but such accuracy gains required nearly 100 labels instances per person in [13][24]. The work [25] also demonstrated that accuracies of person-specific stress detection models, using physiological data, can be improved by about 10% if additional 100 labels per person are available. Unfortunately, requiring so big number of labels per person is unrealistic approach.

Table 1 summarises long-term real-life studies that we found. As we believe that real life studies should last long enough to include sufficient variety of human tasks and conditions, we included in Table 1 only studies that lasted at least four weeks and involved at least four persons. Table 1 shows that during 2013 – 2016 there were five studies [11][12][13][14][17], which experimented with more than 30000 days of data, whereas six studies [6][15][18][20][21][25], performed after 2017, experimented with less than 15 000 days of data. Probably, later studies were affected by COVID, but nevertheless, the trend does not feel right. Table 1 also shows that only half of the studies, performed during 2018-2022, did not use fully supervised AI methods, which means that importance of reducing user labelling effort is not yet fully acknowledged.

2.3 Studies into reducing the need in self-reports

Despite impracticality of fully supervised learning of behavioural models, not so many studies into stress detection aimed at finding solutions to this problem. A good illustration of this situation is a recent review [26] that focused only on supervised methods.

The majority of the studies into reducing the need of self-reports of the target user explored the possibility to use data of similar persons [14][17]. As these works also employed fully supervised training, they required obtaining large sets of labelled data of multiple non-target individuals, and also labelled data of a target individual, although in smaller quantities than person-specific training would require. It was found, however, that success of using data of other individuals depends on similarity between these individuals and the target person: if they are similar enough, accuracies may be by 5- 12% higher than accuracies of general models [14][17], but use of data of not-so-similar persons may result in lower accuracies [17].

Recently, Taylor et al. [20] employed more modern method to exploit similarity between human beings: multi-task learning, which is a type of transfer learning. In this approach deep network is trained so that some parameters are shared between all tasks (a human can be considered a task), whereas other parameters are task specific. This approach requires to collect data from many persons: in [20] data of 104 persons were used in training, and they were clustered based on pre-study questionnaires, such as Big Five features and gender. Hence multi-task learning is a promising approach for cases when data of many persons can be collected and shared, but the drawbacks are data sharing (it may hinder user acceptance because of potential privacy threats) and the need to answer questionnaires, required for user clustering. It would be beneficial to study the ways to cluster people based on their sensor data instead.

Table 1: Summary of long-term real life stress detection studies to date

Ref	Study duration	Number of subjects	Sensors	AI approach
[11]	4 months	35	Phones and wearables	Supervised
[12]	6.5 months	117	Phones and weather	Supervised
[13]	6 weeks	28	Phones	Supervised
[14]	8 weeks	30	Phones	Supervised
[17]	8 weeks	30	Phones	Transfer learning
[15]	4 weeks	25	Phones	Supervised
[6]	10 months	4	Depth cameras	Unsupervised
[21]	4 weeks	65	Phones and wearables	Unsupervised
[20]	1 month	104	Phones and wearables	Multi-task learning
[25]	1-3 months	14	Wearables	Supervised
[18]	8 weeks + 10 weeks	30 + 48	Phones	Supervised

Unsupervised stress detection, to the best of our knowledge, was studied only in a few long-term real-life studies: Tervonen et al. [21] used self-organising maps and Vildjiounaite et al. [6][19] used Hidden Markov Models. Unfortunately, unsupervised methods are difficult to control: they are much more likely to produce completely wrong results than supervised methods, and they are far less accurate: Tervonen et al. [21] reported only 60% accuracy of stress detection on daily basis; the work [6] reported 67% accuracy, but the proposed method was tested on long-term data of a few persons only and hence it is unknown how well it can work for larger population. Semi-supervised methods would be a good compromise because they would allow more control than unsupervised AI methods while requiring fewer labels than supervised methods, but we are not aware of any stress detection studies, employing semi-supervised methods.

3 CHALLENGES AND FUTURE DIRECTIONS FOR HDT LEARNING IN REAL LIFE

Figure 1 presents the main building blocks of AI-based stress detectors. They are created, and their parameters are adjusted, during system training. These main blocks are as follows:

- ***Data***: sensor data and ground truth data, used to train and test the system.
- ***Pre-processor***: a module performing data filtering (e.g., removing corrupted data), data segmentation (e.g., in fixed time windows of a certain length or splits data into segments that have some meaning: for example, separates gestures from not moving hands)
- ***Feature extractor***: a module that extracts from data segments selected features (e.g., maximum value of data in a segment) and, if needed, normalises them.
- ***Feature selector***: a module that selects "best" features from all extracted features. This module is not needed if the set of features is determined and fixed during system development. It is needed, however, if the "best for each user" feature set is determined during system operation separately for each user, or the system removes on features with low variation in their values
- ***Classifier***: a module that takes selected features as an input and outputs its conclusions in a form of a class or a score (for brevity, we will not distinguish between classification and regression models here).
- ***Fusion***: a module that combines data from different sources (e.g., different sensors or different persons). Fusion can be done on feature level: features from different sources are combined and used by the classifier in the same way as features from a single source. This approach requires availability of features from all sources in each data segment. Fusion can be also done on classifier level: in this case, outputs of different classifiers (in a form of scores or decisions) are combined, for example, scores can be summed up or used as an input to another classifier; decisions can be combined by voting. This approach does not require availability of features from all data sources in each data segment, but often, its accuracy is lower than that of feature-level fusion.

AI systems can also employ a single deep learning network to perform feature extraction, feature selection and classification, but training of such networks usually requires large sets of labelled data, which are difficult to collect in real life. Below we will describe challenges and future research directions, related to learning human models in real life.

3.1 Overall system design

System design should consider the following aspects:

1. ***Choice of sensors***: a trade-off between system accuracy and privacy. Usually, the more accurate information the sensor can provide, the more privacy-threatening it can be. For example, video cameras allow to obtain facial expressions and to instantly detect human surprise or anger, but they are more privacy-threatening than depth cameras because currently, depth cameras cannot recognise people. Depth cameras, however, cannot recognise human mood as well as video cameras. Another example: human confusion can be recognised from interactions with the software, for example, if a user tries the same commands repeatedly. If the user listens music, however, repeating the same commands may mean that the user wants to listen the favourite song again. Hence interaction

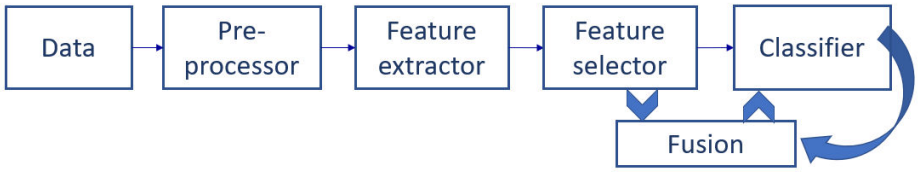

Figure 1: Overview of AI systems

analysis would be more reliable if the system would know which software and which commands the person uses, but this knowledge increases privacy risks. In addition, accuracy depends on required granularity of results because aggregation of data over time can increase accuracy [6].

Therefore, choice of sensors (and provided information details) should depend on user acceptance and on application requirements, such as minimum acceptable accuracy and granularity of human condition assessment. Users should "opt in" for using the system, and should be able to choose which sensors to use, if to use any. For example, the users may in the beginning opt-in for using the most privacy-safe sensors, then become curious about other sensors and start using them too, if they will develop trust in the system – that the system will not harm their work situation.

2. ***How to process data along timeline***. Data usually arrives as a stream, which is split by a pre-processor into segments. Each time segment can then be treated as an independent data instance, i.e., classified separately from other data segments. Such instant classification results can be later combined along timeline. An alternative way is to use sequential classifiers, i.e., methods that process consecutive data instances together. Such methods can, for example, learn that stress state does not instantly transform into a calm state, and use this knowledge to classify a state after stress also as stress, even if instant classification would classify it as a calm state. Which methods would work better, instant or sequential, remains to be studied because the majority of the current approaches employ instant classifiers. The work [6] employed sequential classifier (Hidden Markov Models), but did not compare it with instant classifiers. The work [18] compared sequential classifier LSTM (long short-term memory network) with instant classifiers (Decision Tree and AdaBoost) and observed that instant classifiers achieved higher accuracies in the majority of cases, but data size was probably too small for LSTM.

3. ***How to process data of multiple persons.*** Usually, data of many test subjects are collected, so it is possible either to learn model of each human being independently or to use data of many persons together. In the former case, i.e., learning of person-specific models, the very important choice is, how to minimise number of self-reports, used in system training. For example, is it more beneficial to use a limited number of self-reports for feature selection or for classifier training? To the best of our knowledge, studies into this problem are not abundant, and not concerned with the problem of limited availability of labels. Furthermore, most probably studies with one data type cannot be generalised to other data types because no AI method performs well on any data.

In the latter case, adaptation to each person can be done via transfer learning, and the system developers should choose, whether to use all available data together or to cluster test subjects and to train different cluster models independently on each other. Apparently, clusters should reflect similarity between different human beings, and how to find similar persons, is an open research question. The work [20] clustered people by personality (Big Five factors) and genders, but apparently this is rather coarse clustering, and it requires people to answer questionnaires. Ideally, clustering should be based on sensor data.

To summarise, systematic real-life studies are needed to address the following research problems:

- What kind of ***accuracy/ granularity/ privacy trade-offs*** can be achieved with different sensors and how different sensors can be used together. This requires multi-sensor data collections, which to date were mainly performed in the labs, while the majority of long-term real life studies employed only one sensor, as Table 1 shows.
- How to perform ***mental conditions assessment continuously***: whether sequential or instant classifiers suit better for this task and how past human conditions affect present conditions. To date, use of "past" data for stress detection was rarely studied; most often each time period was assessed independently on the previous periods.
- How to ***minimise number of self-reports***, required from each end user for system adaptation to each individual: for example, whether it is more beneficial to use the self-reports for feature selection of for tuning of classifier parameters. Another approach is to use ***data of multiple persons together***, and this approach requires to study ***acceptability*** of data (or features) sharing and development of ***similarity metrics*** between different individuals

Below we will describe other aspects of the building blocks, which need to be addressed for real life learning of human models.

3.2 Datasets

As our information search shows, many current stress detection studies are still performed in the labs, but stress and calm conditions in the labs differ from that in real life. In lab studies stresses are usually induced by giving test subjects relatively short tasks that are preceded by a relaxation phase; hence, lab conditions do not well represent long strenuous days at work. Non-stressed conditions in a lab may also notably differ from normal office work: for example, in [27] non-stressed state of test subjects was induced by showing them movies of swans and ducks in a lake.

A good example of difficulty to transfer results of lab studies to real life is stress detection on the basis of physiological data. Studies, conducted in the labs, report very high accuracies, e.g.,

combination of EMG (electromyography) and ECG (electrocardiogram) data allowed to achieve 97-100% accuracies in classifying two-three stress classes [28]. On the other hand, a large-scale (1002 subjects) study with test subjects, monitored in the course of their normal lives during five days, reported less optimistic results [29] for three-class stress classification: data of 43% of subjects appeared unsuitable for stress detection. In leave-one-subject-out tests on the remaining dataset a fully supervised classifier (Random Forest) achieved an average F-score 0.43, which is not a high accuracy. The main reasons for so notable difference between lab and real life results are influence of human activities (motion, food intake etc.) on physiological parameters and motion-induced sensor shifts.

The most interesting result of Smets et al. [29], however, is finding that the highest classification accuracies (F-scores above 0.66) were achieved for the test subjects who did not experience stress very often: on average, they reported 86% of "no stress", 12% of "light stress" and 2% of "high stress" cases. On the other hand, F-scores for the most stress-prone subjects (who reported 26% of "no stress", 45% of "light stress" and 29% of "high stress" conditions on average) were below 0.33, i.e., performance as good as random. This means that stress detection on the basis of physiological data may be not the best method for the most high-risk group of people, who experience stress often. Therefore, researchers should aim at collecting real life data with different sensor types.

Long-term data collections in real life require test subjects to regularly provide self-reports, as self-reports are the most common method to obtain data labels in such studies. Unfortunately, human beings usually fail to provide large numbers of self-reports: for example, in 10-day study [30] test subjects on average provided 28% of self-reports they were prompted to provide, and in five-day study of Smets et al. [29] (supplementary material) test subjects provided 42% of self-reports they were expected to provide. Therefore long-term data collections need methods to motivate test subjects to provide self-report and not to drop off. In many studies, the subjects are given a small reward in the end, but continuous motivation during study duration could be more useful. A potential candidate could be gamification, but we are not aware of any studies, employing gamification for data collections in workplaces, where self-reporting should take very little time.

To summarise, the major shortcomings of lab data collections, which complicate transfer of lab results to real life, are the following:

- ***Types of human conditions***. In the labs usually only short-term conditions, such as acute stress, are studied, although long-lasting stress of low intensity may have equal or greater impact on health as a short-term high stress [31]. In real life, stressors include also long-lasting problems, for example, lack of support from colleagues or boring repetitive tasks, which cannot be studied in the labs. Most probably such types of long-lasting stressors cannot be detected using any single sensor. Therefore, the studies should be designed to last long enough to include long-lasting problems, and also to employ different sensor types, to find out which human conditions and which stressors can be detected by which sensors.
- ***Sensor data quality***. Lab studies are often designed to obtain high quality sensor data, which is impossible to achieve in real life. For example, in the labs test subjects are usually positioned so that cameras have frontal view on their faces or whole bodies, and the distance to the cameras is fixed. In real life, on the other hand, humans turn and move, and furniture may obstruct view. For example, in a lab study [24] depth cameras were recording data of the whole body, while in real life study [6] only head trajectories were collected because office furniture obstructed the view on the bodies. Another example: stress detection studies, analysing computer usage data, recorded many data details, e.g., differentiated between letters, numbers, space key, shift key etc., whereas in real work collecting detailed data would rise privacy and security concerns, and high level of details can compromise passwords.
- ***Labels quality***. In lab studies distinction between stress and non-stress conditions is defined by study protocol and is therefore very clear. In addition, labels are well balanced, because study protocols usually place test subjects in stress conditions for about half of the study, and in non-stressed conditions also for about half of the study. In real life, labels are usually self-reports of the subjects and are often missing, while available labels are noisy, because the subjects may hesitate what to report and/ or make mistakes in reporting. Labels are also often imbalanced because some subjects might be almost never stressed, whereas others may be almost always stressed. Durations of different human states (e.g., stress vs. non-stress) in real life studies also vary a lot, and it is not known when transitions between the states occur because test subjects do not provide self-reports each time their conditions change. In long studies, labels can be obtained only one-two times per day, otherwise the subjects would drop off from the studies. Problems with labels lead to a situation when researchers are aware of real life data availability, but chose not to use it. For example, recently a real life data of nurses at work was published [32], but the work [33] decided against using it in the experiments because the authors "found a substantial number of marked sections that exceeded the expected duration of the event, with a further number of short events with no potential for a cooling down period to separate the perceived stressful event from a subject baseline (non-stressed)".

3.3 Data pre-processing

As described in the previous section, in the lab studies distinction between stress and non-stress conditions is clear, and labels are assumed to be trustworthy because they are determined by study protocols. In real life, clear cuts between human states are missing because data contains human states of unknown duration and transitions between states of unknown duration. Labels are usually collected periodically, and in case of stress label it is unclear whether stress occurred in the beginning of reporting period or in the end, or maybe even during previous reporting period (real life study [6] reported comments of the test subjects, which suggest that tiredness during previous day may cause stress next day even if the next day is nothing special by itself). Labels can be also wrong, and very often, labels are missing.

A typical data pre-processing approach is to segment data into chunks of a fixed length: for example, each label can be predicted using data, collected a few hours before the label (while other day data can be ignored), or the whole day is split in time windows of a fixed length. Most often, each day is processed independently on the previous day. The problem of noisy labels is often addressed by employing noise-robust classifiers, e.g., SVM, but these classifiers remain noise robust only if they get not so small numbers of labels.

Pre-processing would benefit from the following studies:

- **How to segment data into meaningful chunks** and to ignore unclear transitions between human states because including such transitions reduces ability of data features to discriminate between important states. For example, "meaningful" may mean a chunk of a continuous work on the same application, or any continuous activity, or using data of the work week instead of processing each day separately. Detection of "continuous activity" may require multi-modal data; hence, it would be beneficial also to study, which data types would be most helpful for data segmentation (these data types can be then used as auxiliary, i.e., only for segmentation, or also for stress detection).
- **How to obtain meaningful labels without requesting end users to provide many labels**. Although large number of labels should help to ignore unclear transitions in the data and to increase noise robustness of classifiers, from practical point of view it is infeasible to request a lot of labels from end users. Hence stress detection would benefit from developing novel approaches to assess trustworthiness of labels. It would be feasible also to develop novel approaches to active learning, for example, to detect frequent behavioural patterns and to ask for labels for these patterns. (A typical active learning would require to ask a user for a label when classifier confidence in its result is low, but great variety of human activities may result in low confidence due to a new activity, not necessarily due to stress).

3.4 Feature extraction and selection

Real life stress detection would require addressing the following research problems:

- **Privacy-safe features:** currently, lab studies use for stress detection features which would be impossible to collect in real life, e.g., computer data logs storing all typed characters, phone calls data, logs of facial data or facial expressions etc. Privacy protection requires finding privacy-safe data features and storing only them.
- **Features robust to changes in real life conditions**. Video analysis is a well-established research area, and features, robust to light and face orientation changes, are already known, but occlusions remain a problem. Computer data analysis is a very young research area, with mainly lab studies. With the current trend of hybrid work, people can use different screens, keyboards and mice at work and at home, but we are not aware of any studies into the problem of using different types of keyboards on different days because this problem never occurs in the lab studies. Analysis of depth camera data is also a relatively young research area, which lab studies analysing data of the whole body [24], whereas in real life occlusions only allow to track head locations [6].
- **Person-adaptive feature selection**: Numerous studies have shown that different people react at stress differently: for example, some people freeze, others fidget [24]. Reactions may also depend on stress reasons, for example, some tasks may require intense interactions with computers, whereas other tasks may require communications with humans. Therefore methods to choose for each person features, most indicative of his/ her problems, are necessary to learn models, suitable for large varieties of personalities and tasks, but current feature selection methods usually require large number of labels and do not perform well when number of labels is small.

To the best of our knowledge, the majority of stress detection studies to date employed handcrafted features. An interesting feature extraction/ selection alternative could be autoencoders. Autoencoders are privacy-safe because they can extract features from the data of each person separately without the need in his/ her labels, and humans cannot easily interpret the extracted features. Whether autoencoders can provide any meaningful features, and whether they can help with adaptation to specifics of each individual or only for dimensionality reduction, however, remains to be studied.

3.5 Classifiers

Realistic real life stress detectors should not require many labels from each end user because users would not provide them. Approaches to achieve this goal fall into two major groups, and none of them is mature enough. First approach is to use in training data of multiple users, where users can be all available users (in this case the resulting model is called general model) or users, most similar to the target person. Maxhuni et al. [17] trained a supervised classifier (Decision Tree) using a mixture of data of a target user and users, most similar to the target user. This approach required about 60 labels of each target person, however, which is not so small number. Multi-task learning [20] also uses data of multiple persons in training of a neural network. Hence these approaches require data sharing between different persons, which is not perfect from privacy viewpoint.

The second approach would be semi-supervised learning of person-specific models. To date, the majority of person-specific stress detectors were trained in fully supervised ways, so semi-supervised stress detection is even less mature than the first approach.

Advanced AI methods, such as deep neural networks and transformers, can integrate three major steps: feature extraction, feature selection and classification, but they typically require many labels for training and hence were not used for human behaviour analysis to date.

From privacy viewpoint, learning of a digital twin of each person should be performed on a personal device of this individual and should use only data of this user for training, but in this case AI method should require not so much device memory and computing power, and should not require many self-reports from the user. The following AI approaches may have potential for doing it:

- **transfer learning approaches, requiring only data of the target person**: for example, to pre-train a general model, using data of big number of persons, and then move this model to the personal device and to fine-tune it using a small number of labels of the target person.
- **cascaded classification**: for example, first AI method, trained on data of multiple persons, is employed to recognise facial expressions, or to extract heart rate, or certain behavioural patterns. Then the outputs of the first stage are adapted to each user by, for example, determining heart rate variability thresholds in person-specific ways (stress usually manifests itself in lower heart rate variability, but how much lower, depends on individuals) or by finding which sequences of facial expressions or behavioural patterns correspond to normal and stressed states of different persons.
- **conventional semi-supervised training of person-specific models:** first, there are intrinsically semi-supervised classifiers, which handle a mixture of labelled and unlabelled data simultaneously (e.g., semi-SVM). Then, there are self-training and co-training methods, which work as follows: first a supervised classifier (or a set of supervised classifiers) is trained using a small set of available labelled data, next the trained classifier(s) classify unlabelled data samples, and the most confident labels are used to re-train the same supervised classifier(s) and so on, until some stopping criteria are met. In case of self-training (when only one classifier participates in the process), most confident labels can be ones located far away from decision boundary. In case of co-training (when several classifiers are employed), different classifiers can be trained on different data modalities or on different subsets of data of the same modality. Usually, the most confident labels are the labels on which all or majority of the classifiers agree [34].
- **cascaded semi-supervised learning of person-specific models**: on the first stage, a model is trained in unsupervised way, and after that its parameters are fine-tuned using a limited set of labels of each person. Alternatively, first stage model can be created using expert knowledge (for example, a regression model). To date, however, studies, that used unsupervised learning [6][19][21], did not attempt at studying next step – a possibility of using small number of labels to increase accuracies of unsupervised classifiers. Generally, semi-supervised learning remains practically unexplored area.

3.6 Fusion

To date, not so many stress detection studies employed fusion of data from multiple sources. When "multiple data sources" are other human beings, fusion is usually performed on feature level, i.e., classifiers are trained on combined features of multiple persons. The work [17] compared this approach with an alternative approach: first person-specific models of different individuals were trained independently on each other, and then for predicting stress of a target person a weighted ensemble of models of similar individuals was used, where weights reflected degrees of similarity to the target person. This weighted ensemble achieved higher accuracy than classifiers, trained on combined data of multiple persons, which is an interesting funding because different works reported that feature-level fusion can achieve higher accuracies than classifier-level fusion at least when "multiple data sources" are different sensors.

Feature-level fusion remains a common approach also when "multiple data sources" are different sensors: in [12] even weather features were used together with phone data features and personality features, the latter being derived from the questionnaires. As feature-level fusion requires simultaneous availability of data from all sources, this approach may be not universally suitable for real life. Real life studies would benefit from research into the following problems:

- **Classifier-level fusion and its personalisation**. Classifier-level fusion for real life data was studied in [11] and [21], exactly because data from different sources were not available simultaneously. In both works phone data were collected during day times, and physiological data were collected at night. Studied fusion methods were not complicated, however: in [11] fusion was performed as weighted sum of day and night stress scores, and in [21] it was "OR" and "AND" fusion of decisions of classifiers for day and night data. More sophisticated fusion approaches were not tested on real life data to date, to the best of our knowledge. For example, stacked classifiers approach would be to train a classifier using outputs of single data source classifiers; classifier selection approach would be to train several classifiers on different data sources or different features and then to select the best classifier for each user. Weighted sum of scores from different classifiers may be also a feasible approach if weights are personalised. Development of personalised fusion methods would be a feasible future research direction, and it would require sizable datasets.
- **Fusion along timeline**. To the best of our knowledge, the majority of studies treat each day as a standalone entity, independently on other days, even though in real life previous day(s) often affect next day(s). Muaremi et al. [11] used first order low-pass filtering to compute long-term stress score of each day. In this approach, long-term stress score of each day was a sum of two values, first value being a long-term stress score of a previous day. Second value was the difference between instant score of the current day and the long-term stress score of a previous day, multiplied by a filter coefficient to limit the maximum difference between scores of the consecutive days. The work [6] summed up stress detection results of different days to obtain stress score of a month. More studies into long-term stress assessment would be needed in future because employees fall sick or resign because of long-lasting stress.
- **Concept drift detection**. Fusion methods could help to detect when human has evolved, and hence his/her twin should be updated. For example, disagreement between the classifiers, trained on the past data, and classifiers, trained on the most recent data, may be indicative of change. Research into detecting human changes is future direction because to date, no studies reported collecting relevant data.

4 SUMMARY AND CONCLUSIONS

A survey of early stress detection methods, conducted in 2016, concluded that methods, suitable for long term use in real life, did not exist yet [35]. Our survey resulted in a conclusion that although more studies were performed after that and more methods were proposed, there are still many open research problems to be solved to develop methods to learn human digital twins for long-term real-life use.

First, the majority of existing works study short-term stress. As long-lasting stress of low intensity may have equal or greater impact on health as a short-term high stress [31], recognition of just short-term stresses could not suffice for wellbeing monitoring. It was also observed that detection of stress is easier in people that experience stress rarely, than in people who experience stress frequently [29], but solution to this problem is yet to be found. Hence recognition of long-term stress is an important future research direction, including considering influence of the previous day(s) on the next day(s). For this purpose, it is important to collect in future large datasets of real-life data. In addition, larger datasets could allow to employ also modern deep networks, which are currently rarely used in stress detection because of insufficient data sizes, except for video data analysis.

Second, the majority of existing studies are still performed using lab data because it is easier to collect; it is clearly labelled, and labels are equally distributed between different classes. In real life studies, on the other hand, labels are noisy and imbalanced, and it is not clear when stress started and ended, and data includes notable greater variety of human activities. These problems lead to anecdotical situation when researchers are aware of real-life data availability but chose not to use it. Therefore, in future methods to deal with noisy labels and lack of clear transitions between human states should be developed.

Third, a large share of studies to date employed single data source. For better acceptance, end users should have the option to choose which sensors to use. Therefore, more studies to compare different sensors are needed, and more fusion methods to process data, not available simultaneously.

Last but not least, stress perception and manifestation (especially in behavioural data, but also in physiological data) are highly person-dependent, and many works reported that person-specific models achieve notably higher accuracies than "one-fits-all" models. In addition, privacy considerations may hinder data sharing. Considering that human digital twins should evolve together with humans, these problems will emerge again and again during human life. Development of AI methods to learn specifics of each user, requiring neither data sharing nor notable self-reporting efforts, remains an important and largely unsolved research problem. It is also necessary to develop AI methods to recognise when human models should be updated.

We believe that the above-listed challenges and research problems should be addressed, and then learning of human digital twins should allow to support human employees in increasingly more cognitively demanding work in different work domains.

ACKNOWLEDGMENTS

The authors thank ITEA Mad@Work project partners and all test subjects of our previous studies. Mad@Work project was funded in Finland by BusinessFinland, grant number 2991/31/2019. Mad@Work was funded in Portugal by Fundo Europeu de Desenvolvimento Regional (FEDER), COMPETE 2020, grant number POCI-01-0247-FEDER-046168.

REFERENCES

[1] Michie, S., Causes and management of stress at work, Occupational and Environmental Medicine 2002, 59: 67-72
[2] Fanny Vainionpää, Marianne Kinnula, Atte Kinnula, Kari Kuutti, and Simo Hosio. 2022. HCI and Digital Twins – A Critical Look: A Literature Review. In Proceedings of the 25th International Academic Mindtrek Conference (Academic Mindtrek '22). Association for Computing Machinery, New York, NY, USA, 75–88. https://doi.org/10.1145/3569219.3569376
[3] Elizabeth M. Argyle, Adrian Marinescu, Max L. Wilson, Glyn Lawson, Sarah Sharples, Physiological indicators of task demand, fatigue, and cognition in future digital manufacturing environments, International Journal of Human-Computer Studies, Volume 145, 2021, 102522, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2020.102522
[4] Louise Wright & Stuart Davidson, How to tell the difference between a model and a digital twin, Advanced Modeling and Simulation in Engineering Sciences volume 7, Article number: 13 (2020)
[5] Naudet, Yannick, Alexandre Baudet, and Margot Risse. "Human Digital Twin in Industry 4.0: Concept and Preliminary Model." IN4PL. 2021.
[6] Elena Vildjiounaite, Ville Huotari, Johanna Kallio, Vesa Kyllönen, Satu-Marja Mäkelä, Georgy Gimel'farb, Unobtrusive assessment of stress of office workers via analysis of their motion trajectories, Pervasive and Mobile Computing, Volume 58, 2019, 101028, ISSN 1574-1192, https://doi.org/10.1016/j.pmcj.2019.05.009.
[7] Johanna Kallio, Elena Vildjiounaite, Jani Koivusaari, Pauli Räsänen, Heidi Similä, Vesa Kyllönen, Salla Muuraiskangas, Jussi Ronkainen, Jari Rehu, Kaisa Vehmas, Assessment of perceived indoor environmental quality, stress and productivity based on environmental sensor data and personality categorization, Building and Environment, Volume 175, 2020, 106787, ISSN 0360-1323, https://doi.org/10.1016/j.buildenv.2020.106787.
[8] Laura P. Jiménez-Mijangos, Jorge Rodríguez-Arce, Rigoberto Martínez-Méndez & José Javier Reyes-Lagos, Advances and challenges in the detection of academic stress and anxiety in the classroom: A literature review and recommendations, Education and Information Technologies (2022)
[9] Aditi Sharma, Kapil Sharma & Akshi Kumar, Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion, Neural Computing and Applications (2022)
[10] Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., Kumar, S., cStress: towards a gold standard for continuous stress assessment in the mobile environment, ACM International Joint Conference on Pervasive and Ubiquitous Computing 2015
[11] Muaremi, A., Arnrich, B., Tröster, G., Towards Measuring Stress with Smartphones and Wearable Devices During Workday and Sleep, BioNanoScience 2013, Vol. 3, Issue 2, pp 172-183
[12] Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Pentland, A., Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits, Proceedings of the 22nd ACM international conference on Multimedia 2014
[13] Ferdous, R., Osmani, V., Mayora, O., Smartphone app usage as a predictor of perceived stress levels at workplace, 9th International Conference on Pervasive Computing Technologies for Healthcare 2015
[14] Garcia-Ceja, E., Osmani, V., Mayora, O., Automatic Stress Detection in Working Environments from Smartphones' Accelerometer Data: A First Step, IEEE Journal of Biomedical and Health Informatics 2016
[15] Ciman, M. and K. Wac, Individuals' Stress Assessment Using Human-Smartphone Interaction Analysis. IEEE Transactions on Affective Computing, 2018. 9(1): p. 51-65.
[16] H. Gimpel, C. Regal, M. Schmidt, Mystress: unobtrusive smartphone-based stress detection, European Conference on Information Systems, 2015.
[17] Maxhuni, A., Hernandez-Leal, P., Sucar, L.E., Osmani, V., Morales, E.F., Mayora, O., Stress modelling and prediction in presence of scarce data, Journal of Biomedical Informatics, 63 (2016) 344-356
[18] Sergio Muñoz, Carlos Á. Iglesias, Oscar Mayora, Venet Osmani, Prediction of stress levels in the workplace using surrounding stress, Information Processing & Management, Volume 59, Issue 6, 2022 https://doi.org/10.1016/j.ipm.2022.103064.
[19] Vildjiounaite, E., *et al.*, Unobtrusive stress detection on the basis of smartphone usage data. Personal and Ubiquitous Computing, 2018. 22(4): p. 671-688
[20] Taylor, S., *et al.*, Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. IEEE Transactions on Affective Computing, 2020. 11(2)

[21] Tervonen, J., *et al.*, Personalized mental stress detection with self-organizing map: From laboratory to the field. Computers in Biology and Medicine, 2020. 124: p. 103935.
[22] Mara Naegelin, Raphael P. Weibel, Jasmine I. Kerr, Victor R. Schinazi, Roberto La Marca, Florian von Wangenheim, Christoph Hoelscher, Andrea Ferrario, An interpretable machine learning approach to multimodal stress detection in a simulated office environment, Journal of Biomedical Informatics, Volume 139, 2023, 104299, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2023.104299.
[23] Carneiro D., Novais P., Pêgo J.M., Sousa N., Neves J., Using Mouse Dynamics to Assess Stress During Online Exams, InInternational Conference on Hybrid Artificial Intelligence Systems 2015, pp. 345-356.
[24] S. Koldijk, M. A. Neerincx and W. Kraaij, "Detecting Work Stress in Offices by Combining Unobtrusive Sensors," in IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 227-239, 1 April-June 2018, doi: 10.1109/TAFFC.2016.2610975.
[25] Tazarv, A., Labbaf, S., Reich, S.M., Dutt, N., Rahmani, A.M., Levorato, M., Personalized Stress Monitoring using Wearable Sensors in Everyday Settings, EMBC 2021
[26] Samriti Sharma, Gurvinder Singh, Manik Sharma, A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans, Computers in Biology and Medicine, Volume 134, 2021, 104450, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104450.
[27] Sharma, N., Dhall, A., Gedeon T., Goecke, R., Thermal spatio-temporal data for stress recognition, EURASIP Journal on Image and Video Processing 2014:28
[28] Pourmohammadi, S. and A. Maleki, Stress detection using ECG and EMG signals: A comprehensive study. Computer Methods and Programs in Biomedicine, 2020. 193: p. 105482.
[29] Smets E., Velazquez, E. R., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., Cornelis, J., Janssens, O., Van Hoecke, S., Claes, S., Van Diest, I., Van Hoof, Ch., Large-scale wearable data reveal digital phenotypes for daily-life stress detection, npj Digital Medicine volume 1, Article number: 67 (2018)
[30] Adams *et al.* Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild, PervasiveHealth 2014, pp. 72-79
[31] Lamb, S., Kwok, K.C.S., A longitudinal investigation of work environment stressors on the performance and wellbeing of office workers, Applied Ergonomics 52 (2016) pp. 104-111
[32] Hosseini, S.; Gottumukkala, R.; Katragadda, S.; Bhupatiraju, R.T.; Ashkar, Z.; Borst, C.W.; Cochran, K. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. Sci. Data 2022, 9, 255
[33] Vos, G., Trinh, K., Sarnyai, Z., & Azghadi, M. R. (2022). Ensemble Machine Learning Model Trained on a New Synthesized Dataset Generalizes Well for Stress Prediction Using Wearable Devices. arXiv preprint arXiv:2209.15146
[34] van Engelen, J.E., Hoos, H.H. A survey on semi-supervised learning. Mach Learn 109, 373–440 (2020). https://doi.org/10.1007/s10994-019-05855-6
[35] Alberdi, A., Aztiria, A., Basarab, A., Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review, Journal of Biomedical Informatics 59 (2016), 49-75.