ITEA Office
High Tech Campus 69 - 3
5656 AG Eindhoven
The Netherlands
T + 31 88 003 6136
E info@itea3.org
W www.itea3.org

ITEA 3 is a EUREKA strategic ICT cluster programme

Mental Health and Productivity Boosting in the Workplace

D2.3 – Pilot Evaluation

Edited by: Henrique Figueiredo, Médis
Date: 16 October 2022
Version: V1.0
Contributing partners: Elena Vildjiounaite (VTT), Diego Fuentes (Hi-Iberia), Omar Nasir (Helvar), Davor Stjelja (Granlund), Päivi Vanttola (FIOH), Hyun-Suk Kim (ETRI), Simão Pedro Rodrigues Ferreira (School of Health of Porto Polytechnic), Kimmo Häyrinen (Uniqair).

Table of Contents

Project acronyms

SW	Software
HW	Hardware
API	Application Programming Interface
ID	Identifier
JSON	JavaScript Object Notation
JWT	JSON Web Token
CLI	command-line interface
LFS	Large File Storage
TLS	Transport Layer Security
HTTP	Hypertext Transfer Protocol
REST	Representational state transfer
OIDS	Open ID Connect
App	Application
ECG	Electro cardiogram
PPG	Photoplethysmogram
IEQ	Indoor Environmental Quality
IAQ	Indoor Air Quality
BMS	Building Management System
BIM	Building Information Modelling
HVAC	Heating, ventilation, and air conditioning
IoT	Internet of things
PIR	Passive Infrared sensors
JS	JavaScript
OS	Operating System
GDPR	General Data Protection Regulation
AI	Artificial Intelligence

1. Introduction

Mad@Work goal is to develop truly unobtrusive, privacy-safe, appealing solutions, smoothly integrated into work environment and appropriate for long-term use in diverse real-life settings. Thereby, Mad@Work proof-of-concept prototypes were tested in long-term pilots, mainly with knowledge workers and respective HR departments in the partners' workplaces. This document presents evaluation results of the pilots, described in the deliverable D2.1.

Mad@Work is aiming at developing a modular solution, so that different partners develop different sensing solutions and different support tools, and end users can decide, which sensing solution(s) and which support tool(s) they prefer. To create such offer to end users, Mad@Work partners need to answer common research questions.

Since Mad@Work aims at using sensor-based stress detection in the support tools, a major common research question is: what can different sensors achieve in real life:

- How to detect stress with ***reasonable accuracy in a realistic way*** for long term use, which means in ***privacy protecting way, requiring little end user efforts*** and ***not requiring any expensive hardware***
    - what is reasonable time granularity: for example, hourly, daily, weekly, etc.
    - what is a realistic number of stress classes: for example, 2 classes (stress / non-stress) or 3 classes (low, medium, high stress)
    - how much efforts of end users are needed in sensing solution: e.g., number of self-reports, need to manually launch a video app, need to charge and put on the wearable device, etc.
- What else can be recognised (e.g. emotions and stressors) and with which time granularity, numbers of classes and end user efforts
- User acceptance of continuous and on-demand sensing solutions

Mad@Work partners had to answer these questions because, although about ten long-term real-life pilots were conducted before Mad@Work, they mainly utilised mobile phones and wearable devices, and the collected datasets are not publicly available[1]. Furthermore, the majority of these works reported using big number of self-reports per end user in classifier training: for example, behaviour-based stress detectors often required over 100 self-reports per end user. This is unrealistic requirement; hence, Mad@Work partners aimed at developing sensing solutions requiring fewer self-reports. Mad@Work partners also aimed at developing new types of sensing solutions, which were not tested in long-term real life pilots before: video-based and based on analysis of computer usage data. These sensing modalities were chosen because they suit to both remote and in-office work, cost nothing (working computer of the user is the only required hardware) and require little end user efforts.

[1] 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

Another major common research question is: how to support the end users in dealing with workplace stress, including persistent stress:

- what is actually a persistent long-term stress
- which stressors are common in knowledge work
- which information should be and should not be available in organisational tools
- which individual and organisational support can be provided for dealing with acute and persistent stress

Prior studies suggested that often workplace support is necessary to help the employees: for example, change of tasks, change of a team, learning new skills, improvement of indoor environment quality or change of office can reduce or eliminate certain stressors. Some of Mad@Work pilots dealt with indoor air quality; some other pilots focused on providing awareness; while other pilots aimed at developing recommendations for acute or persistent stress.

Mad@Work partners did not try to compare which support tools end users would prefer or whether the end users would prefer awareness or recommendations because such choices depend on user personality; hence, the more choices the users have, the higher the chances that they will find a choice to their tastes. The online survey with over 700 respondents, which results are reported in D2.1, have shown that some people would be more interested in organisational tools, while others in individual tools. The survey has also shown that the respondents prefer more accurate, but less frequent stress detection results over more frequent, but less accurate results. Hence all Mad@Work pilots targeted stress detection accuracies 70-80%, and the evaluated support tools were developed according to this accuracy requirement. Other design requirements (answers to the above-listed questions) were obtained via analysis of state of the art and via online questionnaires and focus group studies at the design stage.

Deliverable D3.2 describes AI methods, used in different sensing solutions; D4.2 and D4.3 present individual and organisational tools, developed in the project based on the collected design requirements, and this deliverable presents results of evaluating the developed proof-of-concept prototypes. Section 2 presents common evaluation criteria; Section 3 presents evaluation results per pilot, and Section 4 presents common metrics, which Mad@Work consortium recommends for real life stress detection and mitigation tools. Last section presents consortium conclusions.

2. Evaluation Criteria

The partners have jointly defined a set of common metrics, to be applied to all pilot projects whenever adequate. Below is the list of defined common metrics for each part of the solutions.

2.1. Metrics for AI models

For all AI models developed, these are the common metrics used to evaluate them:

- Modality (type of input sensor data)
- Data size (on which data the model was trained and evaluated)
- Classes (model output)
- Granularity (duration of assessed time period)
- Ground truth (what kind of data served as labels for model training and evaluation)
- Number of required self-reports per end user (how many self-reports per end user are needed to train / adapt the model to this user, example: leave one subject out models need 0 number of self-reports per end user; person-specific models need more than 0 self-reports per user)
- Accuracy (model evaluation on test data)

2.2. Usability and Acceptance metrics

Although each pilot is individual and each solution is unique, the partners have jointly defined common metrics to assess usability and users' satisfaction with the tools. Since Mad@Work aims at modular solution, the partners evaluated sensing solutions and support tools separately. Not all pilots used the below metrics since not all metrics are relevant to all pilots. However, when possible, each metric was applied to each pilot.

- Number of users in the pilot
- Percentage of users approving the sensing solutions and/or support tools
- Percentage of participants that have answered the self-reports and surveys
- Percentage of users worried about data privacy
- Percentage of users considering the sensing solutions and/or support tools easy to use
- Percentage of users considering relevant the information, provided by support tools (e.g., relevancy of recommendations)

3. Evaluation by Pilot

3.1. Pilot 1 – Stress detection and mitigation in location-independent people working in front of a PC (led by Hi-Iberia)

3.1.1. Pilot Description

This pilot intended to deploy and validate part of the Mad@Work platform with people working in front of a PC regardless of their location. Concretely, the pilot aim was to:

- Deploy and validate a video-based stress detection system, which was complemented with online self-questionnaires and physiological data whenever possible.

- Assess the mental health, concretely stress and emotions, in people working with their PC through clips of video recorded through their webcam, and in combination with other sources like self-questionnaires or physiological data.
- Validate if an individual support tool can mitigate stressful situations by recommending relaxing activities to the people participating in the pilot.
- Analyse the acceptance and feasibility of the stress detection system and the individual support tool belonging to the Mad@Work platform among the people participating in the pilot.

For a precise mental health assessment based on stress and emotions detection, it was needed to collect and analyse different kind of data along all pilot phases, that is:

- **Clips of video**, which were recorded through the webcam of the monitored people's PC once a recording session is accepted and initiated from the Mad@Work Web App.
- **Online self-questionnaires**, which were answered through the Mad@Work Web App by each person. Such self-questionnaires were based on commonly-used questionnaires in laboratory and daily life stress experiments such as:
    - Patient Health Questionnaire-4 (PHQ-4),
    - Perceived Stress Scale (PSS), 10 items (once per month)
    - Stress Self-Rating Scale (SSRS),
    - NASA-TLX,
    - Self-Assessment Manikin and Positive and Negative Affect Schedule (PANAS)
- **Physiological data**, such as ECG, heart rate, respiratory rate, blood pressure, which were collected from a smart bracelet worn by each monitored people.

This pilot was carried out regardless the location, either at home or in HI-Iberia office, with 10 knowledge workers, which were recruited voluntarily from the R&D team and SW Developers team. Such pilot had four different stages:

- 1st stage: Interviews with knowledge workers from the R&D team and SW Developers teams, as well as interviews with people working in the HR department.
- 2nd stage: Data collection with knowledge workers from the R&D team and SW Developers teams, at least clips of video.
- 3rd stage: Initial testing & evaluation of video-based stress detection system as a unimodal mental health assessment as well as the multimodal mental health assessment in combination with self-questionnaires and physiological data.
- 4th stage: Final evaluation of mental health assessment as well as the individual support tool, which was implemented in the Mad@Work Web App.

3.1.2. Summary of pilot results and conclusions

The pilot has been carried out regardless the location, either at home or in HI-Iberia office, with 10 knowledge workers, which have been recruited voluntarily from the R&D team and SW Developers team and have participated along different evaluation phases.

Concretely, the final evaluation for this pilot has been performed by 6 workers, who have been able to verify the correct operation of systems and tools implemented, that is, video-based stress detection system, physiological data-based stress detection system, mental health assessment and recommendation system in combination with self-questionnaires and of course, the individual support tool (the Mad@Work Web App), which shows results coming from the above-mentioned systems.

During this final evaluation, it has been possible to verify that the "Mad@Work" concept is well-accepted by workers as long as their data privacy is respected. Most of them consider that the individual support tool is easy to use, since it does not require too much interaction, but not focused on data visualization and recommendation. Additionally, they are happy that the video-based stress detection system is used on-demand instead of being recording continuously.

Anyway, the most valued functionality by workers is the provision of recommendations, given that they do not know what to do when they are experiencing a stress episode.

On the contrary, the main improvement suggested by all workers is about the data visualization through graphs, since they consider that the tool is providing them too much information causing them some perplexity about themselves. Additionally, weekly survey was not considered too much useful by workers at the beginning of the final evaluation, but they finally understood its purpose after further explanations about how to use the individual support tool.

3.1.3. AI Metrics

Some common metrics were defined to evaluate the different models developed and tested in the different pilots. In this pilot, these were the results:

Modality	Data size	Classes	Granularity	Ground truth	Number of required self-reports per end user	Accuracy
Video-based stress detection system	10 people training per session; 100+ video sessions per training; 6 people evaluating now	Stress: stress, non-stress Emotions: angry, happy, sad, neutral	Daily (on demand)	Self-reports	One per week	Stress: 83% Emotions: 77%
Physiological data-based stress detection system (low cost wearables)	WESAD dataset: 15 people; 6 people evaluating now	Stress, non-stress	Hourly	Self-reports	One per week	Stress: 87%

3.1.4. Usability and Acceptance Metrics

Some common metrics were defined to evaluate how users perceived the tools tested, as well as their engagement. In this pilot, these were the results:

- Number of users in the pilot: 10
- Percentage of users approving the sensing solutions and/or support tools: 66%
- Percentage of participants that have answered the self-reports and surveys: 50%
- Percentage of users worried about data privacy: 100%

- Percentage of users considering the sensing solutions and/or support tools easy to use: 100%
- Percentage of users considering relevant the information, provided by support tools (e.g., relevancy of recommendations): 83%

3.1.5. Other relevant results, specific for this pilot

- Data privacy is a very relevant topic for the workers when it comes to sharing personal data with the company
- Stress episodes usually occur in the morning
- Data visualization should be tailored for each user
- Activity recommendations are more than welcome when a stress episode is blocking the worker's mind

3.2. Pilot 2 – Personalized lighting in indoor work spaces (led by Helvar)

3.2.1. Pilot Description

The main idea of Room Comfort Solution is to investigate the indoor air quality and determine the ideal duration for room occupancy through measuring the co2 density and historical occupancy data. We provide information sheets on how long the room should be ideally occupied as well as how long does it take before the air quality degrades too much. We also list down possible action points that should be taken in the occupant feels tiredness or fatigue.

3.2.2. Summary of pilot results and conclusions

Main Results:

- People do not read information sheets multiple times. If they do read it, it would be at most once. Some people do not read these sheets at all, or only glance at it. A few people would take action to improve the air quality themselves, but others expect someone else to take care of it even though everyone can feel the decrease in air quality.
- The undertaken actions include opening the door, having a short break whilst airing the booth and then returning, continuing the meeting in a different location, and having shorter meetings.
- The air quality decreases around 30–40min to an hour. Symptoms of poor air quality are tiredness, fuzzy head, decrease in work productivity, and air feeling stuffy. If a lot of people have used the booth before it makes air quality worse.
- Meeting lengths of the participants: 11 minutes, 30 minutes, 30 minutes, 1 hour, 1 hour. Usually, the length of the meetings are 30 minutes to one hour.

Main Conclusions:

- The information sheets are a useful tool for spreading awareness, but do not necessarily lead to interventions.
- The information must be concise and clearly communicate the risks and mitigation plans.
- Ideally, an automated system that intervenes to improve air quality is preferred over delegating responsibility to the occupant.

3.2.3. AI Metrics

Some common metrics were defined to evaluate the different models developed and tested in the different pilots. In this pilot, these were the results:

Modality	Data size	Classes	Granularity	Ground truth	Number of required self-reports per end user	Accuracy
CO2	30 days (3 sensors in 3 rooms)	N/A	5 minutes fixed frequency	N/A	One	See note below
Occupancy	30 days (3 sensors in 3 rooms)	N/A	Event-based	N/A	One	See note below

Note: The algorithm is unsupervised in nature. It uses statistical tools to identify relationship between air quality and prolonged occupancy. Model accuracy is therefore defined as the majority percentage of samples where the estimated time-limit conforms to the correlation between CO2 and Occupancy, and the CO2 does not exceed safety levels as defined in the Well Standard.

3.2.4. Usability and Acceptance Metrics

Some common metrics were defined to evaluate how users perceived the tools tested, as well as their engagement. In this pilot, these were the results:

- Number of users in the pilot: 5
- Percentage of users approving the sensing solutions and/or support tools: 20% (4 participants were indifferent, 1 user was interested).
- Percentage of participants that have answered the self-reports and surveys: 100%
- Percentage of users worried about data privacy: N/A because IAG sensors do not identify people
- Percentage of users considering the sensing solutions and/or support tools easy to use

3.2.5. Other relevant metrics, specific for this pilot:

- Percentage of users that took an action based on information: 1 user took action (20%), 1 contemplated action but did not have enough time (20%), 3 ignored the information (60%).
- Percentage of users who perceived a significant degradation in air quality. All 5 users noticed a decrease in air quality (100%)

3.3. Pilot 3 – Stress and performance in location independent office workers (led by FIOH & VTT)

3.3.1. Pilot Description

This pilot intended to develop methods to evaluate mental conditions of people working on PC in offices and during remote work. Aims of the study:

- To develop methods to assess mental conditions of knowledge workers utilizing data from computer, mobile phone and/ or environmental sensors, **with the main focus on detecting long-lasting troubles**. Specifically, we intended to:

1. collect long-term real-life data from the above-mentioned sensors, as such databases do not exist;
2. collect self-reports on stress, work content, productivity and (optionally) stressors;
3. collect highly accurate physiological data as a reference;
4. study correlations between various self-reported factors in long term, e.g., between stress and productivity, or between stress and social factors, as well as correlations with physiological parameters;
5. develop methods to detect long-lasting stress and/or other aspects of mental conditions from the collected sensor data (e.g., satisfaction with work content, own productivity or self-reported stressors);
6. assess accuracy of the developed methods at different time granularity;
7. evaluate concept and designs of continuously running organisational barometer;

- To develop and to evaluate gamification methods for motivating knowledge workers to participate in data collections and organisational barometer, because drop-out of data collections is a common problem for employee engagement surveys, as well as in other domains, e.g., health promotion and crowdsourcing.

Initially, we planned to collect data from environmental sensors in the offices, but due to COVID, we adapted the pilot plan to include sensors, suitable for remote work.

3.3.2. Summary of pilot results and conclusions

The pilot was organized with a significant delay compared to the original plan, because (1) due to COVID we had to change our initial plans and to develop a privacy-safe computer data logger; (2) we needed to undergo security assessment and to obtain security certificate to get the permission to install this computer data logger to the work computers of our pilot subjects. This process, however, brought us invaluable knowledge about employer requirements for continuous monitoring of computer usage data in everyday work. Attracting the test subjects after that was relatively easy. Computer usage data were collected unobtrusively, it was non-stop monitoring. Self-reports were collected via an app running on computers or mobile phones. In addition, FIOH collected highly accurate physiological data (saliva samples and FitBit data) and conducted cognitive tests.

After the data were collected, we worked on developing AI methods to analyse the data and evaluated the pilot and the Org. Barometer via questionnaires and focus group studies. Org. Barometer was evaluated by showing the pilot test subjects (1) their individual summaries and (2) their group data. In addition, we conducted focus group study with HR, and focus group study with line managers will be in the end of October - November.

3.3.3. AI Metrics

Some common metrics were defined to evaluate the different models developed and tested in the different pilots. In this pilot, these were the results:

Modality	Data size	Classes	Granularity	Ground truth	Number of required self-reports per end user	Accuracy

Computer	57 persons, 2-6 months per person	Stress, non-stress	Day	Self-reports	30	72%
Computer	57 persons, 2-6 months per person	Stress, non-stress	Quartile	Self-reports	30	84%
Computer	57 persons, 2-6 months per person	Stress, non-stress	Quartile	Self-reports	20	80%
Computer	57 persons, 2-6 months per person	Work simple, work challenging	Day	Self-reports	30	69%
Computer	57 persons, 2-6 months per person	Work simple, work challenging	Quartile	Self-reports	30	89%
Computer	57 persons, 2-6 months per person	Skills “little” or “average” vs. “a lot” or “need learning”	Quartile	Self-reports	30	75%

3.3.4. Usability and Acceptance Metrics

Some common metrics were defined to evaluate how users perceived the tools tested, as well as their engagement. In this pilot, these were the results:

- Number of users in the pilot: 74 (sufficient data for 57)
- Percentage of users approving the sensing solutions and/or support tools
    - Computer usage monitoring: 86.5%
    - Giving self-reports: 59.5%
- Percentage of participants that have answered the self-reports and surveys: 100%
- Percentage of users worried about data privacy: 23.5% (including a little worried users)
- Percentage of users considering the sensing solutions and/or support tools easy to use: 100%
- Percentage of users considering relevant the information, provided by support tools (e.g., relevancy of recommendations): 88%

3.3.5. Other relevant results, specific for this pilot

3.3.5.1 Data collection notes

One the main goals was to collect long-term real-life data from computers and mobile phones, as these sensors are suitable for hybrid work, and to develop AI methods to recognise stress and stressors from these data using self-reports of the pilot subjects. The data collection setup slightly varied between the subjects: the first group (38 subjects) was asked to provide self-reports 3 times a week, while the subjects, recruited later, were asked to provide self-reports every day. Data collection in the first group lasted 4-6 months, while in the second group – 2-4 months. Since computer data were collected every day in all setups, for the first group we collected more unlabelled data than for other subjects. Although altogether 74 subjects were recruited, we report AI metrics for 57 persons because of three reasons:

- Data of 3 subjects were lost due to database connectivity issues
- Several subjects changed jobs or resigned from their organisations before providing enough data

- Some subjects had so many meetings that less than 30 self-reports were available for the days when enough computer data was available, as long meetings may result in nearly zero computer usage
- Some subjects had holidays, business trips or many meetings during data collection period, so even though they provided enough self-reports, they did not provide enough unlabelled data. And the only AI method that was able to achieve reasonably high reported accuracies, required also enough unlabelled data. Tests are still ongoing, so we do not yet know minimum required quantity of unlabelled data

We succeeded to collect a large dataset of computer usage data and self-reports, but we did not succeed to collect phone data due to two reasons: first, modern mobile phones have significantly stronger security and privacy protection functionality than phones used in earlier stress detection studies. So it appeared very difficult to develop phone data logger which would reliably work in different phone models, as every update was either ruining data collection or requiring the test subjects to grant data collection permissions again, and it was tiring. As only a few test subjects volunteered to provide phone data, and they were not happy with updating the data collecting app after phone updates, phone data collection was abandoned rather soon.

On the other hand, 6 subjects out of 15 asked, volunteered to provide video data using the same app that was used in Pilot 6, which shows that attitudes towards video data collection are rather positive nowadays.

3.3.5.2 Gamification notes

Gamification was also used differently for different subjects: in the first group of 38 subjects, the subjects were shown a story and number of points they gained by providing self-reports, but they received the awards for study participation (movie tickets) in any case. Some other subjects were instead of movie tickets using a gift shop where they choose snacks as rewards for the collected points. Not surprisingly, gamification was more positively evaluated by the latter subject group, whereas some subjects in the former group said that they did not understand why gamification was used at all, as they received awards anyway. But we do not know whether gamification helped to prevent drop-outs, as it would require a separate pilot, dedicated to studying effects of gamification. We did not yet compared numbers of self-reports, provided in these two versions: it will be done in November.

3.3.5.3 More Org. Barometer evaluation notes

According to the results of the feedback survey the Organization Barometer was found visually pleasant and very pleasant (90% of respondents), helpful in understanding the reasons of stress and productivity (70% of respondents) and supportive in identification of reasons for positive or negative emotions (50% of respondents). Here, respondents mean “test subjects of the pilot”.

HR group was shown the Org. Barometer demo and asked, if Org. Barometer is useful for them to plan and assess org. wellbeing. The score is 4 (out of 5).

3.3.5.4 Analysis of physiological data and additional questionnaires

Self-reports for developing and testing AI methods were collected via a dedicated app, and due to long study duration, this app was developed to enable very quick reporting. In addition, during first two weeks of data collection 38 knowledge workers provided information regarding total sleep time, working hours, the whole day energy level (9-point scale), and the whole day stress level (9-point scale). In addition, these test subjects provided highly accurate physiological data (by wearing

Firstbeat device) and saliva samples for 72 hours, and they also completed reaction time and cognitive tests on Fridays (up to six Fridays). Stress-relaxation index was extracted from the knowledge workers' 72-hour heartbeat data (Firstbeat). Salivary cortisol concentrations were analysed from the same period (three samples/day).

The group-level models showed that with some associations between the self-reports and the reference variables, the effect sizes were modest. However, in general, self-reported stress linked with deteriorating reference variables (e.g., shorter sleep time before work and poorer flexibility in cognitive functioning), and self-reported excitement linked with recuperating reference variables (e.g., longer sleep time before work and greater physiological relaxation for 24-hours after awakening). This gives some support of their validity as markers for mood. In contrast, high productivity linked in some parts with deteriorating (e.g., higher overall cortisol level and poorer cognitive flexibility) and with some parts with recuperating (e.g., better cognitive accuracy) reference variables, and indeed, both stress and excitement were associated with high productivity.

Sleep time before work was highlighted as a possible predictor of self-reported mood of the upcoming day, while higher cortisol level predicted both excitement and high productivity, but not stress.

Interestingly, the self-reported stress during work had better correlation with whole-day stress report when the latter was dichotomized to indicate 'at least some stress' compared to when it was dichotomized to indicate 'at least a lot of stress'. This implies that the self-reported stress during work indicated relatively mild level of stress, which may explain why stress during work was not associated with the widely used physiological stress marker cortisol.

3.4. Pilot 4 – Learning facility (led by Granlund)

3.4.1. Pilot Description

Granlund's Digital Twin, developed in this project, consists of a knowledge graph depicting the relationships between various building entities like rooms, Indoor Air Quality (IAQ), and Building Management System (BMS) sensors, as well as HVAC systems. By employing the Digital Twin and the correct ontologies, analytics can be scaled widely. Currently, the situation is that different systems and building components are isolated, and given the uniqueness of each building, scaling analytics is profoundly challenging.

3.4.2. Summary of pilot results and conclusions

Throughout the pilot phase, we identified key issues within the HVAC systems of buildings. Consequently, we've instituted specific measures to preemptively detect these issues. Some of these challenges encompass biased or malfunctioning BMS sensors, manual control overrides, consistently inappropriate control curves that excessively boost ventilation, exceedingly high supply air temperatures, and elevated CO2 concentrations.

Additionally, we have crafted a deep learning algorithm capable of predicting occupancy using IAQ data. We've also implemented checks to ensure buildings operate within the suggested IAQ limits, all while optimizing space and energy efficiency. This was done by connecting occupancy data with the data from IAQ sensors and BMS system.

3.4.3. AI Metrics

Some common metrics were defined to evaluate the different models developed and tested in the different pilots. In this pilot, these were the results:

Modality	Data size	Classes	Granularity	Ground truth	Number of required self-reports per end user	Accuracy
Time series	~12.000 rows	0-1	1-3 minutes	Using camera and presence sensors	0	0.82 (MCC score)

3.4.4. Usability and Acceptance Metrics

Since the solution was made for facility managers, no end users were involved in this pilot. The solution was presented to the different stakeholders in building and energy management and their comments (qualitative feedback) were that this is useful tool, but still needs more development regarding additional functions and making it more efficient to make a business out of it. Currently requires a lot of work to set up and this is something that we will work on in continuation through different future research projects.

3.5. Pilot 5 – Safe to breath (led by UniqAir)

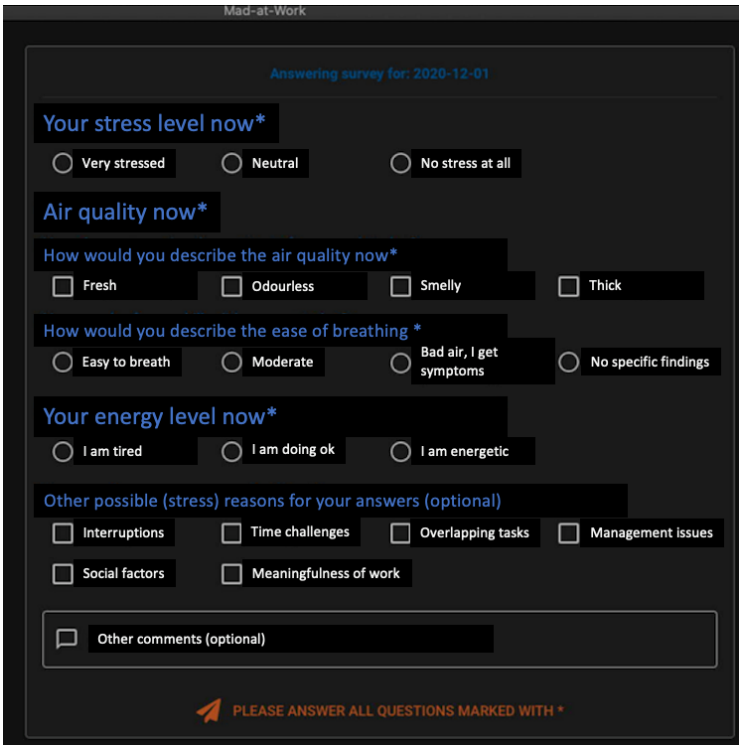
3.5.1. Pilot Description

This pilot monitored workers placing via Empathic Building platform and studied if and how the presence of air purifiers affects the decision regarding which office desk people choose. The pilot also monitored the experienced wellbeing by questionnaire tool. Gathered data was analysed for any correlation between Air Quality, wellbeing experience and presence of purifiers.

3.5.2. Summary of pilot results and conclusions

- Number of users in pilot was small (9) and feedback collecting period was short, 6 weeks.
    - Mostly qualitative over quantitative analysis and findings

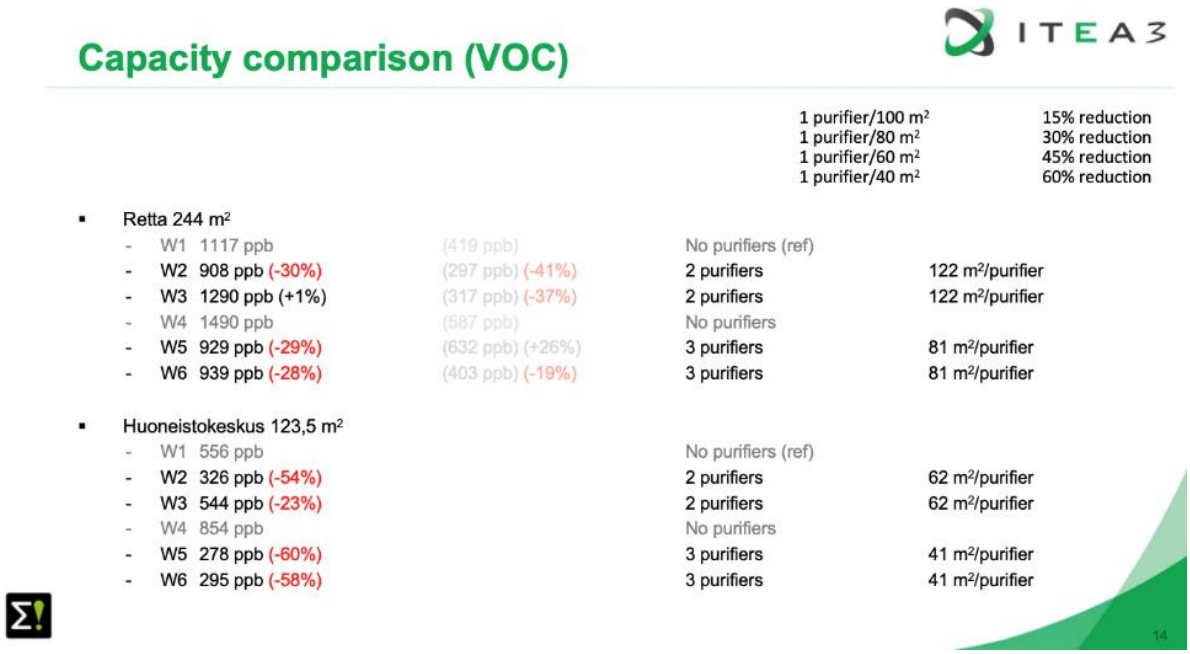
- Human feedback was collected with VTT Survey App



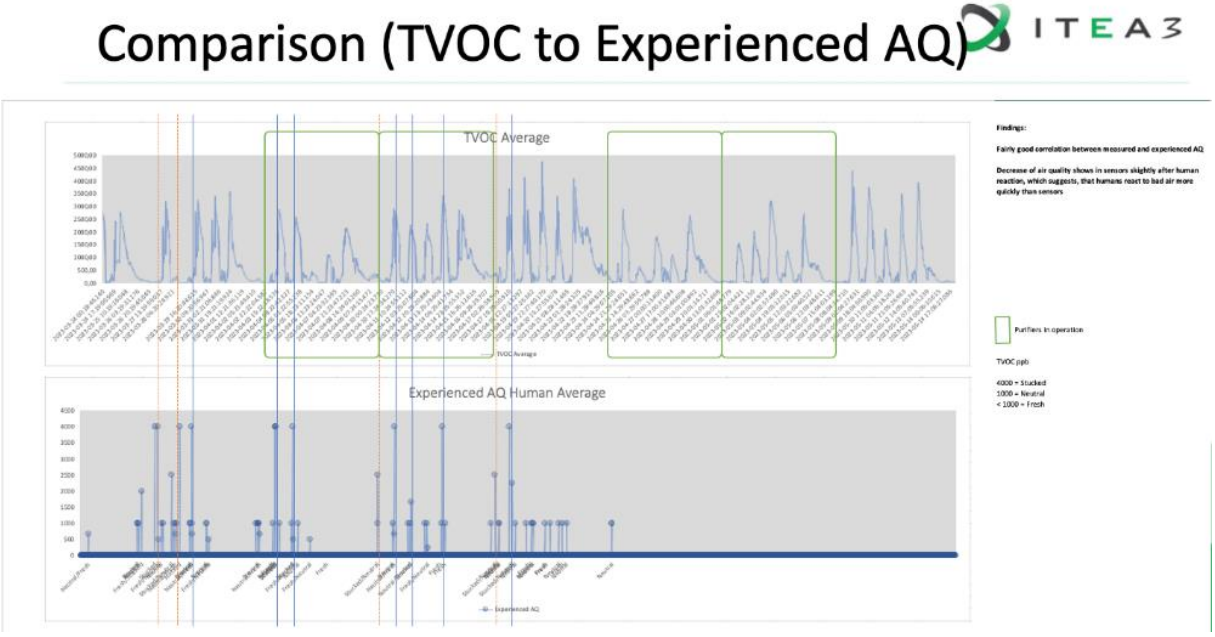
3.5.3. Relevant results, specific for this pilot

CO2, RH and PM stayed at very low level throughout pilot period, some fluctuation in Temperature.

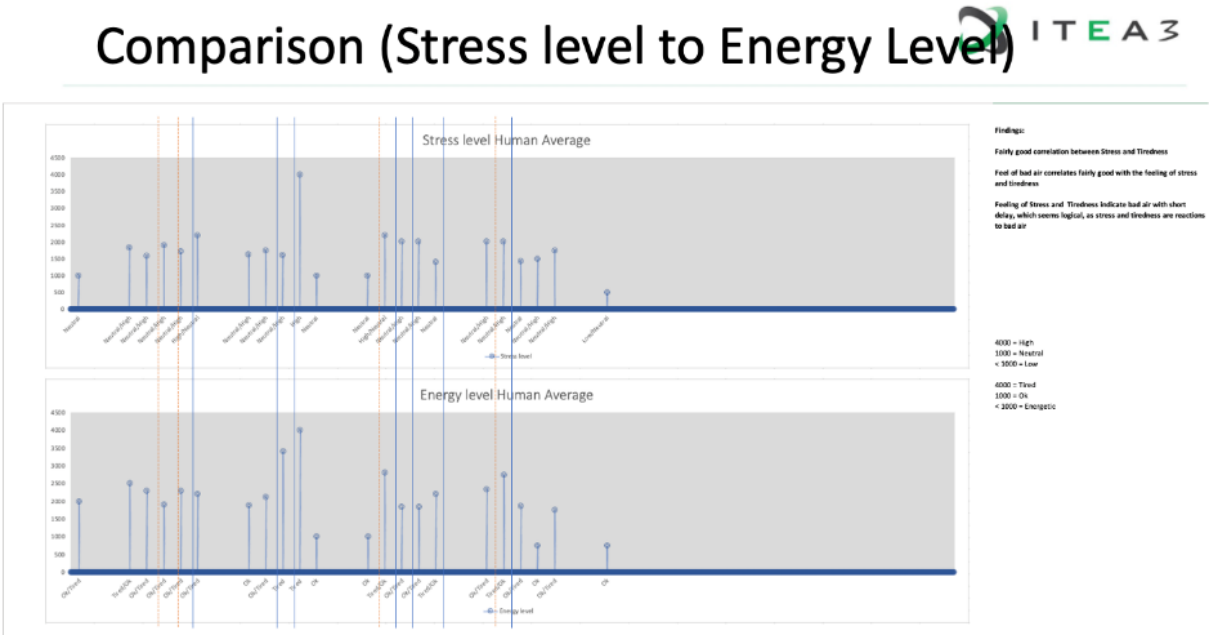
TVOC was the only parameter that was fluctuating wildly and therefore was the main interest and possible reason for reported changes in stress or feelings.



This pilot observed a fairly good correlation between measured and experienced AQ. In pictures the long vertical lines indicate the human reporting of “bad air”



Decrease of air quality shows in sensors slightly after human reaction, which suggests, that humans react to bad air more quickly than sensors.



Comparison of self-reported data and sensor data showed that:

- Feel of bad air correlates fairly well with the feeling of stress and tiredness.
- Feeling of Stress and Tiredness indicates bad air with short delay, which seems logical, as stress and tiredness are reactions to bad air.
- People reported (slight) increase of stress when they reported bad air
- However, the same number of answers reported stress and tiredness during the period when there were no purifiers (and air was poor) and during the period when purifiers were in operation (and air was good). It could be because stress and tiredness are triggered by many other reasons than air quality, and relatively small-scale pilot could not differentiate effects of different stressors. It could be also if stressed persons are more sensitive to air quality and bad air quality adds up to their stress – this issue needs future study.

3.5.4. AI Metrics

Some common metrics were defined to evaluate the different models developed and tested in the different pilots. In this pilot, these metrics were not applicable, since only visual analysis of qualitative results was executed (see pictures and curves above in Section 3.5.3).

3.5.5. Usability and Acceptance Metrics

Some common metrics were defined to evaluate how users perceived the tools tested, as well as their engagement. In this pilot, these were the results:

- Number of users in the pilot: 9
- Percentage of users approving the sensing solutions and/or support tools: Not tracked
- Percentage of participants that have answered the self-reports and surveys:

- 9 users were requested to give feedback daily during working days (at least 1 time per day) during 4 week period. This sums up to 180 individual feedbacks total (20 working days * 9).
- Total of 84 individual feedback was given. This results 84/180 = 47% response rate.

3.6. Pilot 6 – Early detection of stress in the workplace (led by Médis)

3.6.1. Pilot Description

In this pilot, the deployed solution was composed of four main components:

1. Video-based tool, which tracks certain variables through the user's face recognition (perceived emotions, pupil diameter, eye gazing, eye blinking, heart rate variability and facial expressions).
2. Self-assessment questionnaires, which were answered by the user at the end of the period during which he/she was monitored by the video-based tool. The objective is to obtain feedback from the user about how he/she felt during the work period. The questions complement and confirm the data collected by the video-based tool.
3. Self-assessment scales, which were answered by the user to assess mental health disorders (e.g. stress, anxiety, depression) and workload on a monthly basis. The main goal of these scales is to monitor the subject throughout the pilots' period with longer standard/validated scales.
4. In situations in which high stress level is detected, a recommendation system was built– a mental wellbeing support system with scientifically validated recommendations. This tool was developed, however it wasn't piloted due to lack of time.

During the pilot, the videos were be collected or analysed. The data collected was:

- The reading of the variables analysed through face recognition, as stated in the first paragraph
- The calculated stress level for each user/employee periodically
- Users' answers to the self-use questionnaires

The main objectives of this pilot were to:

- Understand if the video-based tool developed has the capability of correctly assessing employees' stress level
- Validate which information should be shared with the employee regarding his/her own stress level, and at what frequency
- Validate which information (aggregated and anonymized) should be shared with employees' managers, and at what frequency
- Identify users' privacy concerns and possible improvements/ changes to the solution to deal with them
- Collect employees' and managers' feedback and inputs regarding the solution deployed, to understand if they would use it and in which circumstances

3.6.2. Summary of pilot results and conclusions

The stress prediction pilot successfully merged technological advancements with machine learning techniques to provide a unique, non-invasive method to detect workplace stress. Utilizing a video-based

plethysmography application, the study collected physiological data from participants during their regular work routines over two months, involving 50 volunteers in each we recovered meaningful data from 28 volunteers. This data was complemented and labeled using a series of questionnaires, gathering information ranging from demographic details to immediate stress perceptions. Through intricate data acquisition and processing methods, including the application of heart rate variability signal features and the use of semi-supervised learning for data labeling, the pilot culminated in the development of several stress detection models. Remarkably, the best-trained model, utilizing the Random Forest algorithm, achieved an accuracy of 86.8% and an F1 score of 87% in predicting binary stress/non-stress outcomes.

The pilot's results underscore the viability of integrating technological tools with machine learning for stress detection, presenting a substantial improvement from traditional methods. The success of the pilot also indicates the potential of physiological data in offering objective insights into employee well-being. This pilot and its approach showcase that a blend of direct feedback from employees, combined with data-driven techniques, can provide a comprehensive, accurate, and meaningful understanding of workplace stress. This paves the way for more proactive and informed organizational interventions in the future.

3.6.3. AI Metrics

Some common metrics were defined to evaluate the different models developed and tested in the different pilots. In this pilot, these were the results:

Modality	Data size	Classes	Granularity	Ground truth	Number of required self-reports per end user	Accuracy
Video	12 Million datapoints downsized to 1291 5-minute blocks	Stress, non-stress	5-minute blocks (real time)	Self-reported stress levels	Minimum 5 work days	86.8% Accuracy 87% F1

3.6.4. Usability and Acceptance Metrics

Some common metrics were defined to evaluate how users perceived the tools tested, as well as their engagement. In this pilot, these were the results:

- Number of users in the pilot: 50 participants (sufficient data for 28)
- Percentage of users approving the sensing solutions and/or support tools
    - 96% of users felt their data was safe
    - 93% felt safe during the pilot
- Percentage of participants that have answered the self-reports and surveys: 62%
- Percentage of users worried about data privacy: 3.5%
- Percentage of users considering the tool easy to use: Not applicable since the pilot was only clicking an icon
- Percentage of users considering the recommendations relevant: Not applicable yet, since the recommendation tool wasn't tested yet. We are actively preparing for the next evaluation, but we were not able to proceed to the recommendation's evaluation on in a real live pilot yet because the first pilot started with long delay and because we need to optimize efficiency and effectiveness in data collection first (see Data collection and recruitment notes below).

3.6.5. Other relevant information specific for this pilot

3.6.5.1 Data collection and recruitment

In our pilot, we initially enrolled 50 participants who voluntarily responded to the outreach email within the company. A significant effort was invested in the onboarding process for each participant, encompassing individual meetings to install the necessary software and to ensure their comfort and confidence with the data we would be collecting. Furthermore, we established a primary communication channel with each participant, facilitated by the investigator responsible for data collection.

However, while the enrollment was promising, there were unforeseen challenges that influenced the final data we were able to obtain. Our pilot coincided with the period of COVID-19 restrictions, a time characterized by a surge in online meetings using platforms like Zoom, Google Meet, and Teams. Given that our solution is video-based, it posed an unexpected challenge. Computers, by design, can handle only one camera request at a time. This meant our participants faced the additional task of toggling the data collection software on and off amidst their already busy online meeting schedules.

Further complicating matters, some of our participants, specifically those in outside sales roles, encountered functional incompatibilities with the webcam-based data collection. Moreover, a segment of our participants reported issues linked to the computing power of their work computers, indicating that their machines struggled to function effectively with our application running in tandem.

While the challenges were multifaceted, our dropout rate was commendably low, with only one participant leaving, who was promptly replaced. Nevertheless, the culmination of these challenges meant that, of the original 50 participants, we were able to obtain comprehensive data from only 28. This outcome, it's essential to note, was not reflective of the participants' lack of interest/commitment but rather a confluence of unforeseen external factors that affected the data collection process.

To address these problems, we've implemented key enhancements, such as automating the data collection software. Now, there's no need for manual toggling; it will activate automatically whenever the webcam is available for use.

3.6.5.2 Delays in the pilot

Our pilot presented us with numerous lessons that have highlighted the importance of foresight, collaboration, and adaptability. At the beginning, we were confronted with unexpected delays during the Data Protection Officer's (DPO) meticulous review. This underscored the necessity to engage with the DPO early, giving them ample time to thoroughly evaluate the project dealing with sensitive data. Similarly, ethical considerations cannot be understated. This calls for a proactive approach—the need to carefully prepare all relevant documentation and justifications upfront for smoother system deployment in real workplaces.

The IT department's involvement was another facet we realized cannot be relegated to the sidelines. Their insights, especially regarding system integrations or infrastructure setups, can have a pronounced effect on system deployment timelines. It is also necessary to reserve time to address various to address unforeseen hitches which can emerge during system installation and integration.

Lastly, the imperativeness of ensuring robust security protocols led us to employ an external firm for penetration testing. While this was non-negotiable, the time invested in the testing and subsequent iterations, if vulnerabilities are identified, is a crucial aspect to be factored into future plans.

These lessons culminate in a series of recommendations for future use of the sensing systems in real offices:

- Initiating processes and seeking necessary approvals well in advance can prevent last-minute bottlenecks.
- Always earmark buffer periods in project schedules to account for unexpected delays, ensuring timelines remain feasible.
- Prioritize transparent communication, ensuring all stakeholders are consistently aligned and apprised of developments.
- Comprehensive documentation from the get-go can considerably streamline processes, especially during review stages

3.7. Pilot 7 – Pilot *WellMind* (led by ETRI4)

3.7.1. Pilot Description

ETRI conducted two pilots. To collect the HR and PPI data from the wearable device and transmit the information to the Galaxy Tablet, the WellMind Application (App) was installed on the Samsung Watch3.

In the first pilot, the WellMindSpace (hereinafter referred to as WSpace) app was developed to collect data from the device in a laboratory environment. The WSpace application possesses a labeling function that permits the annotation of stressful and relaxing task data as stress and non-stress labels, respectively. The WSpace application also featured survey functions for gathering subjective stress questionnaire responses.

Figure 1 Pilot test using the WSpace application

In the second pilot, we developed the WellMindFriend (hereinafter referred to as WFriend) app, which collects data in everyday work environments.

Figure 2 Pilot test using the WFriend application

3.7.2. Summary of pilot results and conclusions

In the first pilot, for the stress task, subjects had to perform a web search on a requested topic and respond via email. Also, subjects had to memorize the answer and then announce it. For relaxation, three types of relaxation tasks were performed alternately. (Close your eyes and relax, stretch, use a massage chair to relax) Data from this pilot test was used in the stress classification model. In this pilot test, 80 participants were involved in the experiment. However, data from 17 subjects were excluded from our analysis for various reasons, including the loss of experimental data, input errors in data labelling, and abnormal data collection caused by the loss of Bluetooth communication with the experimental equipment for a certain period. By applying the Linear Regression algorithm, we achieved 84.9% accuracy.

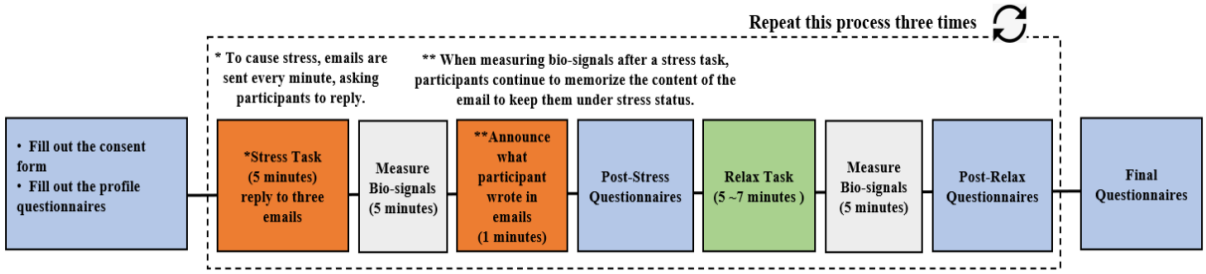


Figure 3 Pilot test procedure.

In the second pilot, to measure the effects of stress and well-being in the office environment at work, 40 people in Group A and 40 in Group B participated. All subjects conducted the pilot test from Monday to Friday. Group A performs HR and PPI data measurements and surveys in the office in response to periodic PUSH alarms four times a day. In Group B, HR, PPI data, and questionnaire data were measured in the morning when subjects arrived at the office. The subject then performs the wellness solution three times a day when he or she wants to rest. The wellbeing solution used resting with eyes closed, stretching, and meditation while watching videos provided by the WFriend app.

In Group A, there were many cases where there were vacation or business trips, so in Group B, this was restricted when recruiting subjects. Therefore, only subjects who had consecutive data on Mondays, Tuesdays, Wednesdays, Thursdays, and Fridays were used for data analysis. In group A, 21 out of 40 experimenters were analysed, and in group B, 35 people were analysed.
As a result of the independent sample t-test on the survey information VAS (Visual Analogue Scale), the VAS value of group A was 39.87 on average, and the VAS value of group B was 40.17 on average, showing similar characteristics. The PSI (Perceived Stress inventory) value of group A was found to be 21.64 on average, and the PSI value of group B was found to be 16.82 on average, showing a significant difference (p = 0.00). In the case of Group B, who took a break while using the wellbeing solution, it can be seen that the subjective stress state decreased. Additionally, it was confirmed that the PSI survey, which consists of 9 questions, expresses stress conditions better than the VAS.

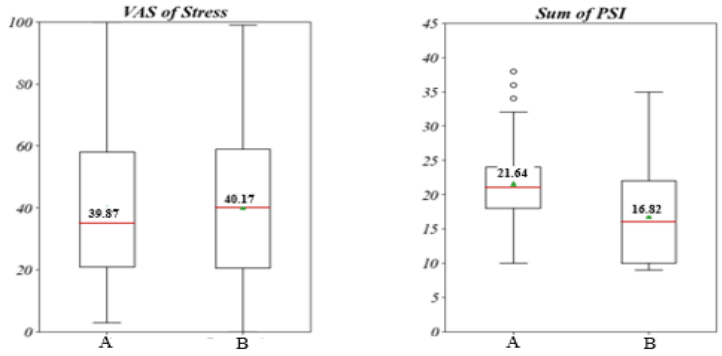


Figure 2 Independent sample t-test results for survey information (VAS/PSI)

3.7.3. AI Metrics

Some common metrics were defined to evaluate the different models developed and tested in the different pilots. In this pilot, these were the results:

Modality	Data size	Classes	Granularity	Ground truth	Number of required self-reports per end user	Accuracy
Physiological data	63 subjects (1.944.984 lines)	Stress, non-stress	Real-time stress detection	Experimental protocol	Conduct a questionnaire when measuring physiological data	84.9%

3.7.4. Usability and acceptance Metrics

Some common metrics were defined to evaluate how users perceived the tools tested, as well as their engagement. In this pilot, these were the results:

- Number of users in the pilot: 40
- Percentage of users approving the sensing solutions and/or support tools: In second pilot study, Group B, who received the wellness solution, conducted an online satisfaction survey on the WFriend app, the mental management system used in the experiment, for five days after the pilot test ended, and analysed the results.
    - 52.5% responded that it was helpful in improving their daily lives and work performance
    - 37.5% responded that it was average
    - 10% said it didn't help reduce their stress

- Percentage of participants that have answered the self-reports and surveys: 100%
- Percentage of users worried about data privacy: This wasn't measured
- Percentage of users considering the sensing solutions and/or support tools easy to use: In second pilot study, Group B, who received the wellness solution, conducted an online satisfaction survey on the WFriend app, the mental management system used in the experiment, for five days after the pilot test ended, and analyzed the results. According to the results of question No. 3* of the System Usability Scale:
    - 25% responded that they would use this system often
    - 55% responded that they would usually use this system
    - 20% responded that they would not use it often
- Percentage of users considering relevant the information, provided by support tools (e.g., relevancy of recommendations): This wasn't measured separately, see above

3.8. Pilot 8 – Indoor Environmental Monitoring – through Air Quality Index (led by BEIA)

3.8.1 Pilot Description

BEIA facilitated the indoor air quality monitoring use case to offer an increased awareness regarding the proposed methods to evaluate mental conditions of people working on PC in offices and during remote work. The case study was organized in one of the working spaces at BEIA location in Bucharest, Romania. The office space is designed to host at any time 10 to 20 workers. In order to collect the air quality measurements, a Libelium Plug and Sense Smart Environment PRO station was used, equipped with the following sensors:

- Temperature and humidity sensor
- Particulate Matter optical sensor (monitoring concentrations of PM10, PM2.5, and PM1)
- CO2 sensor
- O2 sensor
- Luminosity sensor

In addition, self-reports and computer usage data were collected using VTT apps and were used in VTT data analysis.

3.8.2 Summary of pilot results and conclusions

As stated above, computer usage data and self-reports from this pilot were used in VTT data analysis work and contributed to the results, presented in Section 3.3.
Relations between air quality and stress/ productivity is not yet analysed; it is planned for November-December-January. Hence the next session presents only results of evaluating quality of collected data and data visualisations.

3.8.3 Technical Evaluation

The technical evaluation metrics considered for the case study focused on the sensors' data consistency aspects. The following methodology was applied:

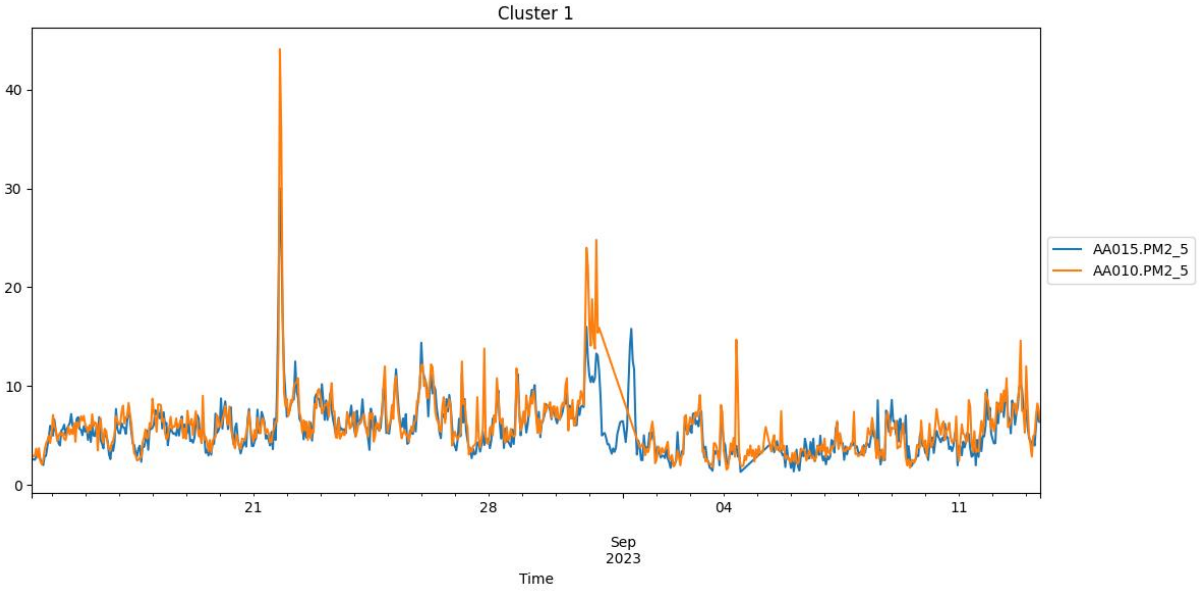
- To ensure precise and consistent monitoring, 10 air quality stations were installed at a singular location.

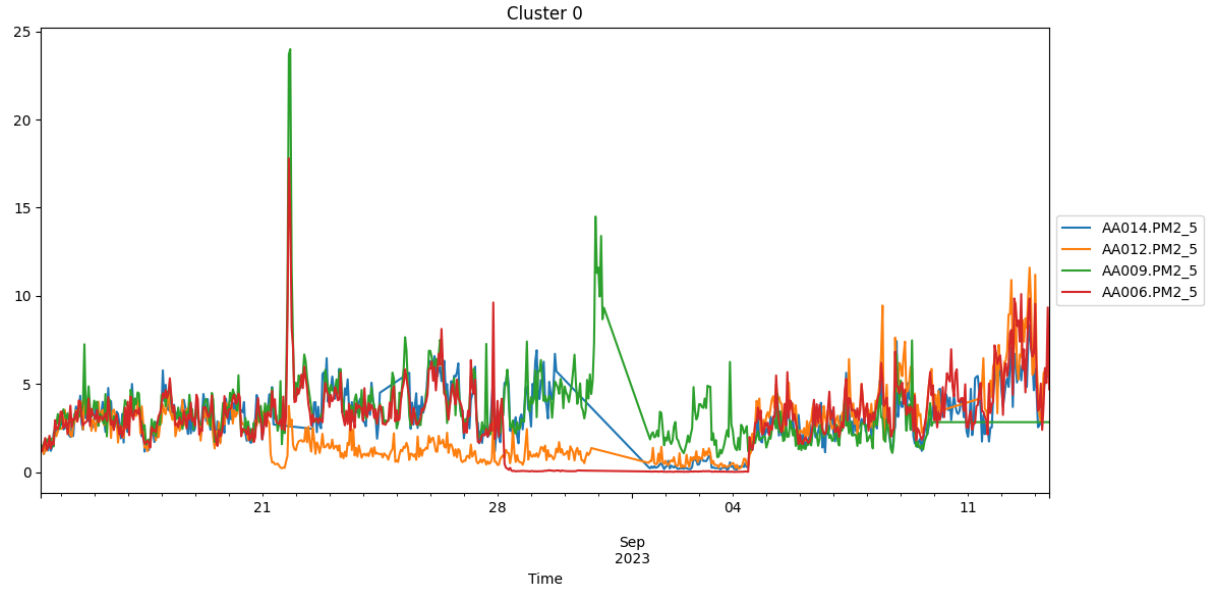
- These stations continuously recorded PM2.5 concentrations, and the data was subsequently exported for a comprehensive analysis.
- Data Preparation
- Load and explore PM2.5 data from various monitoring stations.
- Address missing data via linear interpolation and mean imputation.
- Quantify and visualize the impact of the data cleaning process.
- Initial Data Exploration
- Visualize time-series of PM2.5 measurements post-cleaning.
- Derive descriptive statistics to grasp data trends and variations.

	AA015.PM2_5	AA014.PM2_5	AA013.PM2_5	AA012.PM2_5	AA011.PM2_5
count	721.000000	721.000000	721.000000	721.000000	721.000000
mean	5.560902	3.081791	4.148190	2.393840	4.143682
std	2.675675	1.577478	2.470868	1.808203	2.443308
min	1.320000	0.125000	1.130000	0.193000	1.100000
25%	3.820000	2.100000	3.090000	0.993000	3.070000
50%	5.180000	2.940000	3.866667	2.020000	3.860000
75%	6.730000	4.120000	4.740000	3.410000	4.780000
max	30.000000	8.350000	51.300000	11.600000	51.200000

	AA010.PM2_5	AA009.PM2_5	AA008.PM2_5	AA007.PM2_5	AA006.PM2_5
count	721.000000	721.000000	721.000000	721.000000	721.000000
mean	6.089098	3.477910	4.361311	4.283394	2.791156
std	3.388152	1.968548	1.957809	2.300868	2.130033
min	1.550000	0.843000	1.060000	-0.740000	0.006670
25%	4.050000	2.440000	3.120000	2.760000	1.390000
50%	5.580000	2.930000	4.161029	3.900000	2.880000
75%	7.140000	4.120000	5.220000	5.470000	3.890000
max	44.100000	24.000000	30.400000	29.100000	17.800000

- Statistical Analysis
  - Implement ANOVA to probe differences in PM2.5 levels across stations.
  - Utilize insights to discern potential groupings in the data.
- Advanced Data Exploration
  - Execute cluster analysis, categorizing stations into coherent groups.
  - Visualize and interpret cluster outcomes in the context of PM2.5 levels and geographical distribution.





4. Common Metrics, Recommended for Use in Real-Life Stress Detection and Mitigation Tools

Common metrics, used in this project and recommended for use in other real life stress detection and mitigation tools, can be placed in two groups: first group presents recommendations for sensing solutions, and second group presents recommendations for support tools.

4.1 Recommendations for sensing solutions

- Recognise ***two classes*** of stress and stressors. It does not make much sense to detect stress on finer scale because (1) real life data are challenging, and trying to detect more classes will increase the error and (2) prior research stated that long-lasting stress of low

intensity is not less dangerous than short stress of high intensity, could be even more dangerous. Hence when it comes to stress, it may be more important to distinguish between long-lasting and short-term stress than between high and low intensity of stress.

- Adapt ***granularity of detection*** (that is, which period is evaluated: minute, day, months etc…) to sensor type.
    - ***Physiological and video data*** allow to detect "stress now", and for such kinds of data we recommend to detect
        - number of stress episodes per day
        - number of days per week with some stress episode
    - ***Behavioural data*** do not allow to detect instant stress because unusual behaviour can happen due to many reasons, e.g., meetings. Furthermore, often, stress shows itself in behavioural data not on the stressful day, but on the day after that. Hence classification accuracy for behavioural data often increases when data are aggregated over longer time periods.
- Assess ***prolonged stress*** for **1 month** and **3 months** periods. Based on the WHO's diagnostic tool, the ICSD-11 (ICD-11 (who.int), prolonged stress is something that lasts for at least 1 month. Many existing stress questionnaires ask the respondents to assess their conditions during last 3 months. Hence prolonged stress should be detected with sufficiently high accuracy at least for 3 months period, but 1 month can be a trigger for offering support tools. When using self-reports for detecting prolonged stress, it should be at least ***2 stress reports per week, which would be 29% of the reports***.
- ***Report*** for sensing solutions the following:
    - Accuracy and limitations of the classifier (e.g., if it was trained on data of young managers, it may be not applicable to data of old managers or secretaries)
    - Accuracy reporting should consider that stress data are often imbalanced: some people can be almost never stressed, while others can be almost always stressed. Therefore, it is needed to report not only overall accuracy, but also percents of false negatives and false positives, or F1 score, or AUC (area under curve) measure as these measures allow to better understand the model's ability to distinguish between the two classes

4.2 Recommendations for support tools

As was validated by prior studies and through our focus group studies and online surveys, ***workplace stressors*** include too simple or too challenging work tasks, time pressure, social problems, managerial problems, indoor environment problems, interruptions etc. Some of these stressors can be detected using sensors, for example, air quality. Our study demonstrated that computer usage data analysis allows to differentiate between simple and challenging task, too, which was not attempted by prior studies. Some other stressors, however, are difficult to detect using sensors, for example, bad management. Nevertheless, we suggest that both individual and organisational tools should address all these well-known stressors.

For organisational tools, we suggest that it is useful to recognise the users who can recognize negative impacts of different types of stressors on wellbeing, because it is important also to target users who are unable to do so. To this end, it is important to ***measure the percentage of users who take action***, e.g., follows recommendations because this allows to recognize the effectiveness of the tools.

We would like also to note that in our tools, organisational state is an aggregation of individual stress information. Hence organisational and individual tools reflect the influence of the same stressors. However, individual tools primarily concentrate on identifying and mitigating specific behavioral patterns at the individual level, they adopt a bottom-up approach, commencing with individual needs. On the other hand, organizational tools operate on a grander scale, affecting teams, providers, the entire organization, and even external metrics.

Moreover, individual tools tailor interventions to specific aspects of an individual's behavior in pursuit of stress reduction. In contrast, organizational tools embrace a general strategy, implementing measures that influence various facets of the organization, including its culture, policies, and practices.

5. Conclusions

This document describes the results of evaluating Mad@Work proof-of-concept prototypes in real life pilots and lessons learned when the consortium organized the pilots. These lessons are invaluable for organizing next pilots and for deployment of commercial solutions. For example, since we did not want to use expensive hardware in addition to workplace computers, we were requested to get security clearances to install our SW in workplace computers. We also learned how to explain the potential users how their privacy is protected, which is crucial for gaining user acceptance.

Regarding the evaluation results, due to notable delays in start of the piloting (due to COVID and to above-mentioned security requirements), we were not able to pilot the complete stress detection + support tools systems in next pilots. Some partners will do it later; some partners have evaluated tools by showing to the pilot participants visualizations of their data, which is a valid approach because the evaluated tools anyway show long-term data aggregation.

We would like to note, however, that the most novel and most crucial for user acceptance aspect of Mad@Work is sensor-based detection of stress and stressors, and evaluation of this aspect suggests that the Mad@Work consortium succeeded. First, such detection is achieved via sensor-based monitoring of employees in real work, and the most important evaluation result is, such monitoring was well accepted by the pilot participants. Nobody was “significantly worried” about their privacy, and monitoring was perceived as “easy” by the majority of the subjects. Second, the consortium achieved targeted stress detection accuracies, which is great, considering how novel and challenging was the task. For example, one of the few studies into reducing number of user efforts, required for training stress detectors[2], reported 71% accuracy of three-class (low, medium, high) stress detection from behavioral data (mobile phones), but they used 60 self-reports per each user to train the classifier (many other studies used even greater number of self-reports). Requirement of 60 self-reports means that each user has to provide self-reports for 3 working months before the system starts working, it is quite long time period. In contrast, our method to analyse behavioural data needs 20-30 self-reports per person to achieve the same accuracy in two-class stress detection. And it is the first real life study into stress detection from computer usage data. We are not aware of any real life pilots with video data, either, so we cannot compare our accuracies with state of the art for the same sensors.

As all new technologies, our monitoring solutions would rely on early technology adopters, but reasonably high stress detection accuracies and user acceptance of monitoring suggest that in the

[2] 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

most challenging tasks of the project we achieved success, even though we have not yet finished data analysis tests.

Regarding organizational and individual support tools, we derive organizational stress as aggregation of individual stress. Individual tools primarily concentrate on identifying and mitigating specific behavioral patterns at the individual level, and they adopt a bottom-up approach, commencing with individual needs. Individual tools tailor interventions to specific aspects of an individual's behavior in pursuit of stress reduction, and we were not yet able to evaluate how well we can do this tailoring. There are known methods to achieve tailoring, however, such as tracking of implicit and explicit user feedback.

Organizational tools operate on a grander scale, affecting teams, line managers, HR, and the entire organization. They should influence various facets of the organization, including its culture, policies, and practices. Such tools can be only evaluated in pilots, lasting several years, so we could not do it in this project. We have developed and successfully evaluated visualisations of org. barometer and indoor environmental control tools, however.

Our findings highlight the necessity of a synergistic approach, where both individual and organizational tools play pivotal roles. Individual tools enhance personal resilience and well-being, facilitating a bottom-up approach to stress reduction. In parallel, organizational tools operate on a broader canvas, impacting diverse organizational dimensions, with potential implications for overall performance and external relationships.

Recognizing and connecting these approaches is paramount. By doing so, organizations can develop a comprehensive and effective stress management strategy that not only benefits individual well-being but also contributes to organizational success and external stakeholder relations.

From the corporate point of view, there is a potential for the creation of new value propositions and business models directed at non-invasive stress detection, dynamic data collection and objective stress analysis. These services could be sold separately or as part of consultancy/insurance offers to corporate clients.