xR4DRAMA

Extended Reality For DisasteR management And Media planning

H2020-952133

D3.4

Stress level detection techniques v1

Dissemination level:	Public
Contractual date of delivery:	Month 13, 30 November 2021
Actual date of delivery:	Month 14, 1 December 2021
Work package:	WP3
Task:	T3.4 Stress level detection
Type:	Demonstrator
Approval Status:	Final version
Version:	1.0
Number of pages:	20
Filename:	D3.4_xR4Drama_StressLevelDetection TechniquesV1_v1.0.pdf

Abstract

This deliverable describes the initial versions and outcomes of the stress level detection component of xR4DRAMA developed in T3.4 of WP3. This component is responsible for developing body sensor-based and/F1 audio signal-based technologies for the assessment of the stress level experienced by actors in a situation. The results of the audio and sensor modules will later be merged to obtain one unique stress prediction.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	13-10-2021	Table of contents	Montserrat Marimon
0.2	19-11-2021	First draft	Montserrat Marimon
0.3	23-11-2021	Draft for Internal review by Martina Monego (AAWA)	Montserrat Marimon
1.0	01-12-2021	Final version	Montserrat Marimon

Author list

Organization	Name	Contact Information
UPF	Montserrat Marimon	montserrat.marimon@upf.edu
UPF	Joan Codina	joan.codina@upf.edu
STX	Maria Pacelli	m.pacelli@smartex.it
CERTH	Stamatis Samaras	sstamatis@iti.gr

Executive Summary

This deliverable describes the initial versions and outcomes of the stress level detection component of xR4DRAMA developed in T3.4 of WP3. This component is responsible for developing body sensor-based and audio signal-based technologies for the assessment of the stress level experienced by actors in a given situation. The results of the audio and sensor modules will later be merged to obtain one unique stress prediction.

Abbreviations and Acronyms

CCC	Concordance Correlation Coefficient
DT	Decision Tree
ECG	Electrocardiogram
EDA	Electrodermal activity
EMG	Electromyogram
FRs	First Responders
GA	Genetic Algorithm
GSR	Galvanic Skin Resistance
HNR	Harmonics-to-Noise Ratio
HR	Heart Rate
IMU	Inertial Measurement Unit
kNN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Machine Learning
MSE	Mean Squared Error
MuSe	Multimodal Sentiment
NHR	Noise-to-Harmonics Ratio
PCA	Principal Component Analysis
RF	Random Forest
RFE	Recursive Feature Elimination
RSP	Respiration
SCWT	Stroop Color and Word Test
SVM	Support Vector Machines
SVR	Super Vector Regression
XGB	eXtreme Gradient Boosting

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

The stress level detection component of xR4DRAMA, developed in T3.4 of WP3, is responsible for developing body sensor-based and audio signal-based technologies for the assessment of the stress level experienced by actors in a situation.

The audio-based stress detection system is designed to work with voice recordings from different sources, in particular phone calls (from citizens to emergency numbers) and voice messages from first responders (FRs). In addition, the stress level of the first responders will also be assessed through physiological signals measured through a smart sensing vest developed to collect electrocardiograph, inertial measurement unit and respiration measurements data. The results of the audio and sensor modules will later be merged to obtain one unique stress prediction.

The position of the stress level detection component in the xR4DRAMA architecture is depicted in Figure 1.

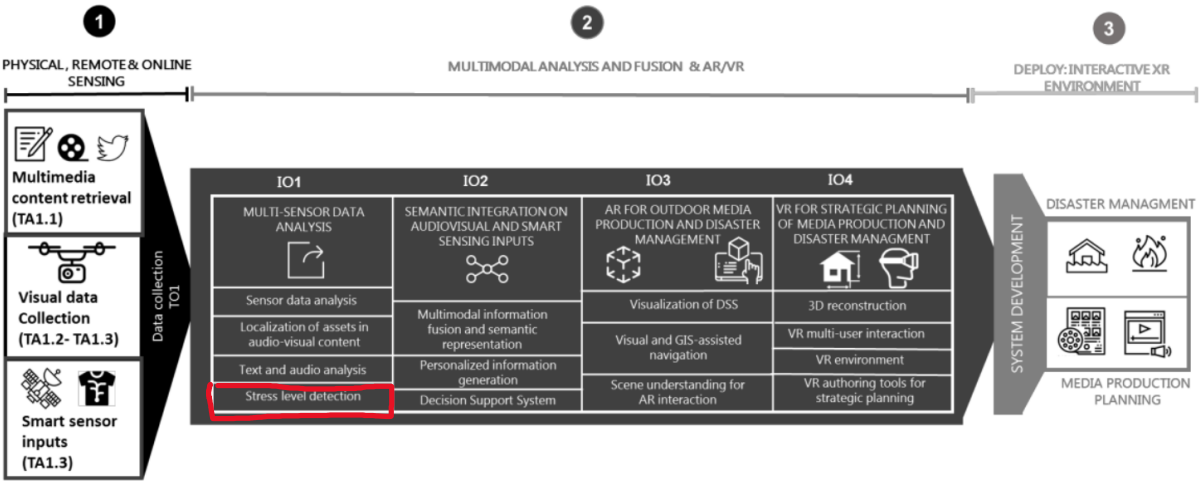

Figure 1: The stress level detection component in the xR4DRAMA architecture.

In this deliverable, we describe the regression model for the analysis of physiological sensors for stress detection, the current baseline system to detect audio-based stress, and the fusion technique that will be deployed in the first prototype for combining the results of the audio and sensor modules regarding stress. We also describe the experiment we have designed to produce our own dataset.

2 BODY-SENSOR BASED TECHNIQUES

In this chapter, the regression analysis of physiological sensors for stress detection is described. The overall task of stress detection also includes classification experiments, which are mentioned in Deliverable 3.1. The regression model is chosen to be deployed in the prototype. First, we present relevant related work and the physiological data acquisition module we have developed for xR4DRAMA.

2.1 Related work

Stress detection by using different modalities has been studied widely recently and is usually based on physiological signals. A well-known publicly available dataset for stress detection from multimodal physiological signals is the WESAD dataset (Schmidt et al, 2018). It consists of wrist- and chest-worn devices containing the following modalities: blood volume pulse, electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RSP), body temperature, and three-axis acceleration. Fifteen subjects have been received interchanged stimuli for amusement and stress conditions. Multiple works have been based on the WESAD dataset for building stress detection systems. In the work of (Bobade and Vani, 2020) multiple machine learning and a simple deep learning method have been assessed for stress detection. Their results show superiority of the deep learning method, which achieved an accuracy of 95.21% for binary classification of stress. In (Indikawati and Winiarti, 2019) only the wrist data were used to build a personalized stress detection system. Three different machine learning methods were used: namely, Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). The RF classifier achieved the higher results, which were 88%-99% accuracy between the subjects. The author in (Siirtola, 2019) conducted experiments to find the optimal set of sensors to build a commercial smartwatch to detect stress levels. In this line, using only the wrist data, they performed feature extraction as described in (Smidt et al, 2018) and performed experiments using different modality combinations. By applying leave-one-out validation, they resulted in 87.4±10.4% accuracy when using Linear Discriminant Analysis (LDA) classifier and the temperature, blood volume pulse and Heart Rate (HR) modalities.

Apart for deploying only physiological signals from wearable devices, other solutions include the fusion of physiological signals along with behavioural data. In the work of (Walambe et al., 2021), a combination of HR variability, skin conductance, camera recordings, body postures, and computer interactions has been proposed. The system aims to detect stress due to workload. The authors proposed two different fusion methods, early and late fusion, both based on neural networks. Results revealed that the early fusion method achieves the best accuracy score up to 96.09%. In (Aigrain et al., 2016) HR, Galvanic Skin Resistance (GSR), EMG, RSP, skin temperature and video modalities were deployed for stress detection during mental arithmetic tasks. In this work three different methods were used to assess the stress levels of the subjects: self-assessment, external assessment from other subjects, and assessment from experts. The system is based on extracting facial, posture, and physiological features and using Support Vector Machines (SVM) classifier, achieving 85.5% F1-score[1]. The

[1] F1-score is a measure to test the accuracy. It is the harmonic mean of precision and recall.

authors in (Giakoumis et al., 2021) used GSR, ECG, accelerometer, body movement, head position, posture, and occurrence of specific gestures to detect stress using the Stroop Color and Word Test (SCWT)[2] as stimuli. Behavioural and physiological features were extracted from the different modalities and, when using LDA classifier, a total of 100% accuracy was achieved.

2.2 Data acquisition

In the xR4DRAMA platform, a physiological data acquisition module is expected with the aim to acquire physiological data to monitor the stress level of First Responders (FRs) in the field for the Disaster management use case. The textile sensing platform architecture (shown in Figure 2) includes the following sensing parts, which are explained in more detail in Deliverable 3.1:

- two textile electrodes to acquire ECG signal;
- one textile respiratory movement sensor (Respiration signal);
- one jack connector to plug the garment to the electronic device;
- one pocket to hold the portable electronic during the activity.

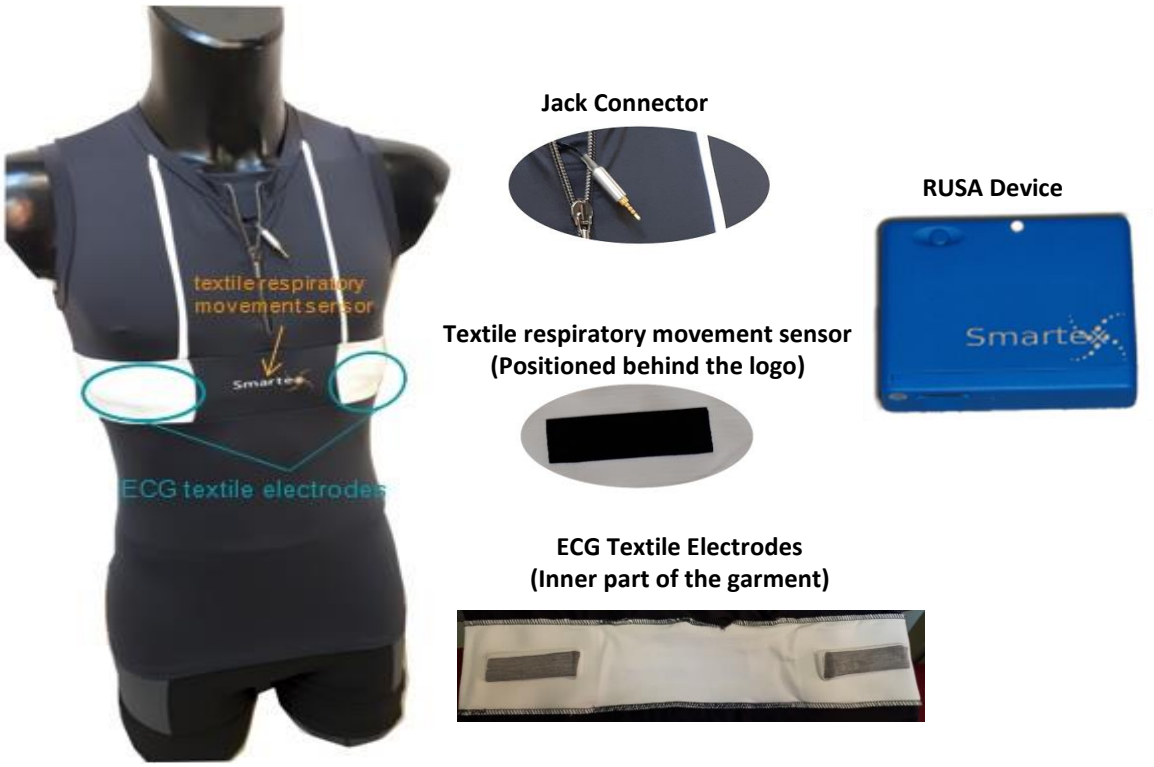

Figure 2: Wearable sensing platform architecture

[2] The tasks that shows the delay in reaction time when there is a mismatch between the name of a color and the color it is printed on.

The wearable system is also able to detect the trunk movements and the posture trough an Inertial platform (Accelerometers, Gyro and IMUs sensors) integrated in the portable electronic device (RUSA). The wireless transmission is achieved by Bluetooth module, that is the WT12, a standard Bluetooth 2.1 module produced by Bluegiga Technologies.

The acquired signals can be analysed separately or combined. For the individual analysis of the sensors, the framework that will be adopted is the following:

- extracting the raw measurements, online or offline, depending on whether the analysis will be in real time or not;
- filtering of the raw signals;
- extracting relevant features. The features depend on the initial signal;
- feeding of the features to a classifier to detect stress or activities.

The sampling rate of raw signals are listed in Table 1.

Table 1: Sampling rate of Raw signals

Name	Sampling rate (Hz)
ECG	250
Respiration	25
Inertial platforms	25

2.3 Data analysis

For the sensor-based stress detection, we used the ECG, RSP, and Inertial Measurement Unit (IMU) sensors. The IMU sensors include the accelerometer, gyroscope, magnetometer, and quaternion sensors. Feature extraction was applied to all these sensors. The features were extracted using a 60-second window with 50% overlap. We used the data of all subjects that were monitored. In total 94 heart rate, 28 respiration rates, and 192 IMU (16 per single-axis data) features were extracted for a total of 314 features. The HR features include statistical and frequency features regarding the signal and the R-R intervals[3], along with HR variability features. The respiration features include statistical and frequency features of the signal, breathing rate, and breath-to-breath intervals. The IMU features include simple statistical and frequency features from the IMU signals.

After extracting the features, the data were split into train and test with an 80/20 ratio. We applied four different ML algorithms; namely SVM, k-Nearest Neighbors (kNN), RF, and eXtreme Gradient Boosting trees (XGB) to perform regression of the stress level since the stress level is a continuous variable. The evaluation was performed by computing the Mean Squared Error (MSE) metric. Before computing the MSE we normalized the values of stress level to be in the range of 0 to 1. We tested each modality alone, all different combinations of modalities in early level fusion and two late level fusion methods: mean and median. Results on the different fusion methods can be seen in Table 2. As we show in the table, the concatenation of all features when using the XGB algorithm has the best result reaching a MSE of 0.073.

[3] The physiological phenomenon of variation in the time interval between heartbeats.

Table 2: Results of early and late fusion methods for stress detection

	HR	RSP	IMU	HR + RSP	HR + IMU	RSP + IMU	HR + RSP + IMU	Late mean	Late median
SVM	0.1709	0.1530	0.1305	0.1723	0.1306	0.1305	0.1305	0.1412	0.1363
kNN	0.1439	0.1553	0.1107	0.1285	0.1106	0.1106	0.1107	0.1170	0.1125
RF	0.1113	0.1280	0.0918	0.1073	0.0916	0.0871	0.0886	0.0984	0.1025
XGB	0.1237	0.1307	0.0844	0.1092	0.0835	0.0858	**0.0730**	0.0958	0.1006

Apart from using fusion methods, we also applied feature selection methods. For feature selection, we used Recursive Feature Elimination (RFE) with RF as the base algorithm, Principal Component Analysis (PCA) with 20 components, and Genetic Algorithm (GA) based feature selection. The results of the different feature selection methods can be seen in Table 3. Results reveal that GA has the best performance when combined with the XGB algorithm achieving a MSE of 0.0567. The best score out of all experiments was the GA with the XGB algorithm.

Table 3: Results of feature selection methods for stress detection

	RFE	**PCA**	**GA**
SVM	0.1052	0.1201	0.1305
kNN	0.1023	0.1106	0.1106
RF	0.0790	0.1044	0.0742
XGB	0.0772	0.0953	**0.0567**

3 AUDIO-SIGNAL BASED TECHNIQUES

Speech is a powerful biosensor; it is produced by a complex combination of physical and cognitive processes.

Despite this richness, voice recording is a non-intrusive and ubiquitous sensor that nowadays is becoming very accessible with the usage of smartphones. Speech contains many levels of information, from voice quality to the content of the speech itself. The analysis of these different information levels (some of them more intrusive than others) can bring indirect monitoring of some physiological and cognitive states of the user. One of these states that can be extracted is the stress level of the user.

"Stress can be defined as the reaction that people may have when they are subject to demands and pressures which do not correspond to their knowledge and abilities and that can challenge their ability to cope. Stress occurs in a wide range of [work] circumstances but is often made worse when employees feel they have little support from supervisors and colleagues, as well as little control over work processes.

Pressure at the workplace is unavoidable … and may even keep workers alert, motivated, able to work and learn.... However, when that pressure becomes excessive or otherwise unmanageable it leads to stress. Stress can damage an employees' health and the business performance."[4]

In the last decade there have been different works that try to extract the stress level from voice. The techniques used have evolved during these years, first because they can use smartphones to capture the audio and second because they are incorporating deep learning techniques.

In xR4DRAMA we follow a similar evolution. First, we start with a baseline based on classical schema to dig into deep learning in the following months of the project.

3.1 System integration

The audio-based stress detection system is designed to work with voice recordings from different sources, in particular phone calls (from citizens to emergency numbers), voice messages from FRs, and possibly other voice recordings from FRs that are recorded through the mobile phone that is used also to collect the sensor data described in Section 2.

The stress detection system works on audio files of variable but finite length (e.g., 30 seconds or longer) and can be called through a REST-like API. For each audio file, it returns a stress level estimation between 0 and 1, along with some metadata (such as the timestamp, etc.).

To use the stress detection system on long audio recordings or continuous audio streams, these will need to be previously split into separate audio files (e.g., using sliding time windows) to obtain a numeric prediction for each of these segments. This is not handled as part of the stress detection system itself and will need to be implemented separately according to the source of the voice recording.

[4] https://www.who.int/news-room/q-a-detail/ccupational-health-stress-at-the-workplace

The results of the audio-based stress detection are then combined with the sensor-based estimations to produce an integrated estimate of the stress level. This combination is performed by a separate fusion component that is also in charge of communicating with the rest of the xR4DRAMA platform to make the results available as needed (as described in Section 4).

3.2 Current baseline system

The current baseline system to detect stress, displayed in Figure 3, is based on a classical Machine Learning (ML) schema with the following steps:

- Speech processing and feature extraction: the speech is processed using a Praat[5] script that extracts a set of acoustic and voice quality features.

  The acoustic features are the following: the frequency of the voice fundamental (F0) and its standard deviation, max value of F0, min F0, range of F0, F0 slope and intensity, ratio of pauses, average pause length, speech rate, articulation rate, average syllable duration, and effective duration of the speech. Rate features were based on the algorithm proposed by de Jong[6] without using transcriptions to keep them as acoustic features measured in a language-independent way (and avoiding ethical issues).

  Voice quality features were also provided by Praat and included all jitter and shimmer available measurements: jitter_loc, jitter_abs, jitter_rap, jitter_ppq5, jitter_ddp, shimmer_loc, shimmer_dB, shimmer_apq3, shimmer_apq5, shimmer_apq11, and shimmer_dda, plus the following harmonicity-based features: harmonicity autocorrelation, noise-to-harmonics ratio (NHR) and harmonics-to-noise ratio (HNR).

- Training: the features extracted can then be used to train a ML model. But to do so, a set of train data is needed. The train data must be composed of some audio recordings with an annotation of the corresponding stress level. Due to privacy issues and as voice can be considered a biomarker, it's difficult to find open train sets available for research. So before having a model a set of train data is needed. Some actions were conducted to obtain this training data.

Figure 3: Current baseline system to detect stress

[5] https://www.fon.hum.uva.nl/praat/

[6] https://link.springer.com/article/10.3758/BRM.41.2.385

3.2.1 Training data

AAWA and DW, the two user partners participating in the project, provided some audio recordings obtained from some real scenarios. These audios were a good starting point to understand and evaluate some of the extra difficulties that must be considered when developing the system, basically: noise (from the environment where the user is talking and audio quality of the devices used to transmit the speech), the kind of messages (duration, content), and speaker variability (professionals, external users). However, this data was not annotated with any stress level indicator, nor did it contain any biological signals (EGG or Breath), so it was insufficient to be used to train or evaluate the system.

As part of the xR4DRAMA project, it was planned to collect training data, in a controlled experiment that would consider all the project needs, which include some physical and psychological stressors, containing all the data that will be available afterwards (speech and body-sensors) and the stress levels (self-reported by the users). The experiment is described in the next section and is currently underway, so the data was not available to train a system.

In April 2021, and as part of the ACM International Multimedia Conference, the MuSe (Multimodal Sentiment) Challenge[7] increased the challenge to include stress detection. From the xR4DRAMA consortium we were aware of this data too late to participate in the challenge, but we could subscribe and get the data. The Muse Dataset is composed of 6 hours of speech annotated with valence and arousal levels, every 0,5 seconds. The stress level is associated with a high value of Arousal (Activated) and an Unpleasant value of Valence (low value).

3.2.2 Baseline system

A first system trained on the Muse dataset and using the classical schema of feature extraction and train a Super Vector Regression (SVR) obtained a Concordance Correlation Coefficient (CCC) of 0.16, much lower than the baseline system of the MuSe Challenge (0.49), which used voice and other physiological measurements like EGG and respiration and BPM signals. Half of the participants did not beat the baseline and all the solutions adopted used deep learning, indicating that this is the way to go (as already stated in the project).

The system (adapted to the MuSe Challenge requirements) produces a prediction of valence and arousal every 0.5 seconds. To do so, it follows the next steps:

1. For each period (0.5 seconds), it extracts the last 10 seconds of sound.
2. These 10 seconds are then analyzed to extract the sound features, using the Praat script, as described above.
3. All the features vectors obtained from all the audio signals, were expanded with the corresponding arousal and valence values.
4. A SVR for Valence and another for Arousal were trained.

From the first analysis performed on this dataset we saw that the 0.5 seconds is a too short period, for two reasons: if we check the desired output, the stress level of the user does not

[7] https://www.muse-challenge.org/

change so quickly, and if we check the inputs, the prosodic features need more than 0.5 seconds to be evaluated. For this reason we used different windows and observed that longer windows (up to 10 seconds) were improving the results. Finally we also observed that each user has very different parameters, and for this reason we think that for known users it can be useful to have a reference signal of their basal state, and then compare the incoming audios with this basal state to better adjust the predictions.

3.3 Training data

As already mentioned, to train any system, data is needed, and at the beginning of the project we observed the need to produce some training data. Once we got knowledge about the MuSe data, we evaluated the need to still produce our own dataset. And we arrived at the conclusion that differences on the experiment design could be crucial in adapting (maybe fine tune) the system to the xR4DRAMA needs.

The MuSe dataset is composed of 69 job interviews. To stress the users, after a brief period of preparation, the subjects are asked to give an oral presentation, within a job-interview setting.

We thought that for the xR4DRAMA use cases the stressing situation should also include some physiological stressors, because first responders are often under psychological and physiological stress. For this reason, the experiment to collect data was design using known stressors for both aspects (physiological and psychological). The stressors selected are:

- Psychological:
    - The colors challenge. The user is presented with some slides with some words written in different colors (Figure 4) and, in a short period of time, must spell out the color in which the words are written. The challenge is that the words are color names, producing a confusion.

Figure 4: Colour change challenge

    - Explain how it has been the day. This is not a stressing challenge, but it is used to get different stress values.
    - Listen to relaxing music.
    - From rom 1324, subtract from 17 to 17. If the user makes a mistake, he/she must start over.
    - Explain a stressful situation in your life.

- Physiological
    - Put a hand in cold water (2º C) for two minutes, make a pause, and put it again.
    - Go upstairs for 4 levels and down again
    - Tie and untie their shoes (after the exercise)

The different challenges were combined so that the user talks while performing some of the physiological challenges. At different moments in time the user is asked to report the stress level he/she feels.

To record the experiment a video was done (translated to Italian) so that all users follow the same sequence with the same timings.

4 FUSION MODULE

The results of the audio and sensor modules regarding stress will later be combined, to receive a more accurate result on the stress level. The fusion technique of averaging predictions was selected as the most appropriate since the audio and sensor signals will not be equal in sizes. The results of the two modalities will be combined in time periods were there were both audio and sensors' input. Let S_x be the predicted stress level based on sensors and A_x be the predicted level of stress based on audio, with the stress levels ranging from 0 to 100. Then the result of fusion for the x case will be:

$$F_x = \frac{S_x + A_x}{2}$$

5 CONCLUSIONS

In this deliverable, we have presented the methods that we have developed for both body sensor-based and audio signal-based stress level detection and the technique to combine them that will be used in the first prototype.

Regarding physiological sensor-based detection, a regression model is chosen to be deployed in the prototype. For voice-based stress detection we start with a baseline based on classical schema to dig into deep learning in the following months of the project. In the next version, we will present this new approach and the results we achieve for stress detection, combining both techniques, using our own dataset.

6 REFERENCES

Aigrain, J., Spodenkiewicz, M., Dubuisson, S., Detyniecki, M., Cohen, D., & Chetouani, M. (2016). Multimodal stress detection from multiple assessments. *IEEE Transactions on Affective Computing*, *9*(4), 491-506.
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7752842

Bobade, P., & Vani, M. (2020, July). Stress detection with machine learning and deep learning using multimodal physiological data. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 51-57). IEEE.
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9183244

Giakoumis D, Drosou A, Cipresso P, Tzovaras D, Hassapis G, Gaggioli A, et al. (2012) Using Activity-Related Behavioural Features towards More Effective Automatic Stress Detection. PLoS ONE 7(9): e43571.
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0043571

Indikawati, F. I., & Winiarti, S. (2020, March). Stress detection from multimodal wearable sensor data. In *IOP Conference Series: Materials Science and Engineering* (Vol. 771, No. 1, p. 012028). IOP Publishing.
https://iopscience.iop.org/article/10.1088/1757-899X/771/1/012028

Siirtola, P. (2019, September). Continuous stress detection using the sensors of commercial smartwatch. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (pp. 1198-1201).
https://dl.acm.org/doi/pdf/10.1145/3341162.3344831

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018, October). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction* (pp. 400-408).
https://doi.org/10.1145/3242969.3242985

Walambe, R., Nayak, P., Bhardwaj, A., & Kotecha, K. (2021). Employing Multimodal Machine Learning for Stress Detection. *Journal of Healthcare Engineering*, *2021*.
https://www.hindawi.com/journals/jhe/2021/9356452/#materials-and-methods