The ForDigitStress Dataset: A Multi-Modal Dataset for Automatic Stress Recognition

Alexander Heimerl, Pooja Prajod, Silvan Mertes, Tobias Baur, Matthias Kraus, Ailin Liu, Helen Risack, Nicolas Rohleder, Elisabeth André, and Linda Becker

Abstract—We present a multi-modal stress dataset that uses digital job interviews to induce stress. The dataset provides multi-modal data of 40 participants including audio, video (motion capturing, facial landmarks, eye tracking) as well as physiological information (photoplethysmography, electrodermal activity). In addition to that, the dataset contains time-continuous annotations for stress and occurred emotions (e.g., shame, anger, anxiety, and surprise). In order to establish a baseline, five different machine learning classifiers (Support Vector Machine, K-Nearest Neighbors, Random Forest, Feed-forward Neural Network, and Long-Short-Term Memory Network) have been trained and evaluated on the presented dataset for a binary stress classification task. The best-performing classifier has been a Long-Short-Term Memory Network, which achieved an accuracy of 91.7% and an F1-score of 90.2%. The ForDigitStress dataset is freely available to other researchers.

Index Terms—Stress, stress dataset, multimodal dataset, digital stress, stress physiology, job interviews, affective computing

1 Introduction

Stress is the body's response to any demand or threat [1]. It is a normal physiological reaction to perceived danger or challenge and can be beneficial in small doses, e.g., stress can support improving performance or memory functions [2] [3]. However, chronic stress can have a negative impact on both physical and mental health. Chronic stress may lead to a variety of mental health problems, including anxiety and depression. It has also the potential to make existing mental health conditions worse. Moreover, stress is able to induce physical symptoms such as headaches, muscle tension, and fatigue [4]. Continuously high exposure to stress increases the risk of heart disease, weakens the immune system, and may promote unhealthy behaviors such as overeating, smoking, and drinking alcohol [4].

Among the many sources of stress, work-related stress is one of the most widespread and often seems inevitable. Work stress is also associated with health problems (e.g., [5], [6], [7]). Therefore, there is a need to understand stressful situations at work and provide coping strategies on how to deal with them in order to prevent chronic stress. However, besides the fact that long-term exposure to work related stressors can be associated with stress, acute situations are equally capable of triggering perceived and biological stress responses. Job interviews have been identified as one of the major stressors in a work-related context for many reasons. They often involve a lot of uncertainty, pressure, and potential rejection. In research, job interview scenarios have recently become a popular use case for studying how to recognize and regulate stress as a result of being a natural stress-inducing event [8], [9], [10].

As remote job interviews have become common practice in response to the restrictions created by the SARS-CoV-2 crisis, such a setting has been used for collecting a novel multi-modal stress data set in a naturalistic setting. For data collection, we recorded signals from various sources including audio, video (motion capturing, facial recognition, eye tracking) and physiological data (photoplethysmography (PPG), electrodermal activity (EDA)). We gathered data from 40 participants who took part in remote interview sessions, resulting in approximately 56 hours of multi-modal data. For data annotation, participants self-reported stressful situations during the interview as well as their perceived emotions. In addition, two experienced psychologists annotated the interviews frame-by-frame using equal stress and emotion labels. Calculating the inter-rater reliability for the individual labels resulted in substantial to almost perfect agreement (Cohen's $\kappa > 0.7$ for all labels). In addition to that, salivary cortisol levels were assessed in order to investigate whether the participants experienced a biological stress response during the interviews.

For automatically classifying the participants' stress levels during the interview, the collected signal information was used to produce a rich high-level feature set. The set contains EDA, heart rate variability (HRV), body key points, facial landmarks including action units, acoustic frequency, and spectral features. Further, a pupil feature set was created based on the latent space features of an autoencoder that has been trained on close-up videos of the eye. In addition to that, the pupil diameter was extracted as well. For reducing the dimensionality of the input feature vector, PCA (Principal Component Analysis) was applied. The classification

- *Alexander Heimerl, Pooja Prajod, Silvan Mertes, Tobias Baur, Matthias Kraus, Ailin Liu, Helen Risack, Elisabeth André: Lab for Human-Centered AI, Augsburg University, 86159 Augsburg, Germany. E-mail: alexander.heimerl@uni-a.de, pooja.prajod@uni-a.de, silvan.mertes@uni-a.de, tobias.baur@uni-a.de, matthias.kraus@uni-a.de, ailin.liu@rwth-aachen.de, helenrisack@yahoo.de, elisabeth.andre@uni-a.de*
- *Nicolas Rohleder: Department of Psychology, Friedrich-Alexander University, 91052 Erlangen, Germany. E-mail: nicolas.rohleder@fau.de*
- *Linda Becker: Professorship for General Psychology, Vinzenz Pallotti University, 56179 Vallendar, Germany. E-mail: linda.becker@vp-uni.de*

task was formulated as a binary stress recognition task (stress vs. no stress). We used and compared the performance of five different machine-learning classifiers - SVM (Support Vector Machine), KNN (K-Nearest Neighbors), NN (Feed-forward Neural Network), RFC (Random Forest Classifier), and LSTM (Long-Short-Term Memory Network). Evaluation of the classifiers revealed that an LSTM approach using all modalities as input led to the best recognition of the participants' stress levels. An LSTM approach also performed best for the majority of modalities. Comparing the different modalities for their impact on the recognition performance, GEMAPS features had the highest accuracy and F_1-scores.

The proposed dataset makes the following contributions to the research community. First, we provide data collected in a realistic stress setting that has been validated by the analysis of salivary cortisol levels in order to assess whether a biological stress response was triggered during the interviews. Second, the dataset was annotated frame-wise using a discrete labeling approach enabling online stress recognition. Third, we provide a multi-modal stress dataset containing established as well as mostly overlooked modalities, e.g., close-up eye features, to provide a promising non-invasive modality for stress detection.

The structure of this article is as follows: In Section 2, we present background and related work regarding existing stress data sets. Section 3 describes the data collection process including design principles, the recording system, properties of the data set, as well as the annotation procedure and feature extraction methods. The method for automatic stress recognition is also explained in detail in Section 3. The results of the performance of the different machine-learning classifiers are presented in Section 4 and discussed in Section 5. Finally, conclusions are provided in Section 6, and ethical considerations are presented in Section 7.

2 Background and Related Work

As of today, multiple stress datasets for the automatic recognition of stress are available. The datasets differ in the modalities used for stress recognition and the stimuli employed to induce stress. Those stimuli range from highly controlled lab settings to realistic real-world scenarios. Table 1 displays an overview of some of the existing stress datasets which are described in the following sections.

2.1 Controlled Laboratory Environments and Stress Tests

The *Trier Social Stress Test* (TSST; [20]) and the Stroop test [21] are standardized methodologies to elicit stress in clinically validated settings. The Stroop test induces cognitive stress through incongruent stimuli, while the Trier test involves interview-style presentations and arithmetic tasks.

The WESAD corpus by Schmidt et al. [19] uses the TSST as a stimulus and provides physiological data, e.g. blood volume pulse (BVP), electrocardiogram (ECG) and electrodermal activity (EDA). Additionally, it features diverse stress-related annotations, including affective states like neutrality, stress, and amusement, obtained from a range of self-report questionnaires.

Similarly, the UBFC-Phys dataset introduced by Sabour et al. [18] used an approach inspired by the TSST to induce stress. While also providing physiological data like BVP and EDA, that dataset contains stress states derived from pulse rate variability and EDA.

For the *Multimodal Dataset for Psychological Stress Detection* (MDPSD) corpus provided by Chen et al. [12], stress was induced using a variety of tests, including the classic Stroop Color-Word Test, the Rotation Letter Test, the Stroop Number-Size Test and the Kraepelin Test. Facial videos, PPG and EDA data are provided. Stress annotations were obtained through self-assessment where participants had to rate their perceived stress on a five-point scale, ranging from no stress to high stress.

2.2 Simulation of Real-World Stressors

To obtain stress-related data in a more realistic context, dedicated efforts have been invested in creating experimental setups that replicate natural conditions where people typically experience stress. These endeavors can be regarded as a balance between controlled laboratory environments and uncontrolled naturalistic settings.

Koldjik et al. [17] introduced the SWELL dataset where they tried to simulate stress-inducing office work by applying time pressure in combination with typical work interruptions like emails. Besides various physiological modalities like heart rate (HR), HRV and EDA, the participants' facial expressions and body posture as well as interaction data were recorded. In order to assess the subjective experience during the study they relied on various validated questionnaires to gather data about task load, mental effort, emotional response, and perceived stress.

Nakashima et al. [14] also aimed to simulate work-related stress. They distinguished between three different states, i.e., "relaxed", "concentrated", and "stressed". The different states were induced by landscape videos for relaxation, Stroop Color-Word test, and Information Pick-Up test. Different test variations were used to induce a state of stress or concentration. After each test, participants filled out the NASA-Task Load Index questionnaire. During the experiment, they recorded the participants' posture by using pressure distribution sensors on the chair and floor. In addition to that, they collected EDA, BVP and HR. Also, apart from us, they were the only ones considering eye-tracking data. However, they focused on different features extracted from the eye-tracker. They collected blinks, fixations, saccades, and scans. Whereas, we mainly focus on pupillometry features.

2.3 Real-World Stressors and Multimodal Data

Moreover, endeavors have been undertaken to gather data in the everyday lives of individuals, often spanning extended durations. These methodologies facilitate a comprehensive understanding of stress dynamics within authentic real-world settings. Nevertheless, they contend with the challenge of handling noisy data that lacks detailed continuous annotations. Additionally, uncertainty exists concerning the nature and timing of stressors.

Healey and Picard [15] presented a dataset for *Stress Recognition in Automobile Drivers* using highly realistic real-world stressors instead of rather controlled approaches to

Name	Number of Participants	Duration per Participant	Stress Stimulus	Modality	Annotation
CLAS [11]	62	30min	Math problems test, Stroop test, Logic problems test	ECG, PPG, EDA, Three-axis Acceleration	Cognitive load, Valence-Arousal
MDPSD [12]	120	<4min	Stroop Color-Word Test, Rotation Letter Test, Stroop Number-Size Test, Kraepelin Test	Facial videos, PPG, EDA	Stress self-assessment
MuSE [13]	28	45min	Final exams period	HR, EDA, Breathing Rate, Skin Temperature, Audio, Video, Thermal Recordings of the Face	Perceived Stress Scale, Self-Assessment Manikins (SAM), Big-5 personality scores
Stress Recognition in Daily Work [14]	10	30min	Stroop Color-Word Test, Information Pick Up Test	Pressure Distribution Sensors, EDA, BVP, HR, Eye Tracking	NASA-Task Load Index
Stress Recognition in Automobile Drivers [15]	24	>50min	Open road driving	ECG, electromyogram (EMG), EDA, Breathing Rate	Free scale stress rating, forced scale stress rating
SWEET study [16]	1002	continuous monitoring for 5 consecutive days	Daily life	ECG, EDA, Skin Temperature, Three-axis Acceleration	Perceived Stress Scale, Pittsburgh Sleep Quality Index, Depression Anxiety Stress Scales, RAND-36 self-reported stress through ecological momentary assessments, Leuven Postprandial Distress Scale, SAM
SWELL [17]	25	3h	Office work with time pressure and email interruptions	HR, HRV, EDA, Facial Expressions, Body Posture, Computer Interaction	NASA-Task Load Index, Rating Scale Mental Effort, SAM, perceived stress on visual analog scale
UBFC-Phys [18]	56	9min	Based on TSST (speech task, arithmetic task)	BVP, EDA, Video	cognitive anxiety, somatic anxiety, self-confidence
WESAD [19]	15	2h	TSST	BVP, ECG, EDA, EMG, Respiration, Body Temperature, Three-axis Acceleration	Three different affective states (neutral, stress, amusement), Positive and Negative Affect Schedule, State-Trait Anxiety Inventory, SAM, Short Stress State Questionnaire, assessment of Stressed, Frustrated, Happy and Sad
ForDigitStress	40	60min	Digital job interview	HR, HRV EDA, BVP, audio, video, body key points, facial landmarks, action units, OpenPose, GEMAPS, Pupillometry features	Time-continuous stress annotations, time-continuous emotion annotations, Coping Inventory for Stressful Situations, State-Trait Anxiety-Depression Inventory, Big-5 personality scores, Perceived Stress Scale

TABLE 1
Overview of existing stress datasets

induce stress. Here, they induced stress by letting the subjects perform open-road drives. Besides recording physiological data, stress annotations were obtained through self-assessment questionnaires using free scale and forced scale stress ratings.

The Multimodal Stressed Emotion (MuSE) dataset introduced by Jaiswal et al. [13] also used a real-world stressor, but in contrast to other datasets, they did not induce stress by simulating a specific scenario themselves. They made use of the final exams period at a university as an external stressor. Therefore, they recruited 28 college students and recorded them in two sessions, one during the finals period and one afterwards. During the recordings, they confronted the participants with various emotional stimuli. Afterwards, the participants self-reported their perceived stress and emotions. Moreover, additional emotion annotations have been created by employing Amazon MTurk workers.

Similar to the MuSE dataset, the Stress in the Work EnvironmEnT (SWEET) study [16] also relied on naturally occurring external stressors as a stimulus. They investigated the participants' perceived stress during their daily lives for five consecutive days. Throughout those five days, they collected physiological data with wearables, contextual information (e.g., location, incoming messages) provided by a smartphone, and self-reported stress. The daily self-assessment was done with a smartphone application that questions the user 12 times a day about their perceived stress.

The CLAS corpus presented by Markova et al. [11] provides valence-arousal labels as well as cognitive load annotations to situations where stress was induced by a math problems test, a Stroop test, and a logic problems test. Additionally, physiological data, such as ECG, PPG and EDA is provided.

Further, datasets exist that are based on non-physiological feature sets. For example, the Dreaddit corpus presented by Turcan et al. [22] contains a collection of social media posts that were annotated regarding stress by Amazon MTurk workers.

2.4 The Need for the ForDigitStress Dataset

Altogether, a large variety of different stress datasets already exist and are available to the research community. However, out of the stress datasets listed in Table 1, only four are freely accessible - WESAD, SWELL, UBFC-Phys, and CLAS. Furthermore, existing datasets show some drawbacks regarding stress labels and recorded modalities, we discuss these in more detail below.

Consequently, there remains a demand for additional freely accessible stress datasets that fill these gaps. With the collection of the ForDigitStress dataset, we decided to develop a setup combining the advantages of laboratory and naturalistic conditions. The TSST test inspired the setup, but we allowed for an interactive scenario where the participant has to engage in a job interview that replicates a real interview as much as possible. Information about how to access the ForDigitStress dataset is provided in subsection 3.6.

When considering provided stress labels, existing stress datasets predominantly are labelled through self-report stress questionnaires or similar assessments. Those approaches come with the disadvantage of being subjective and, more importantly, yielding annotations of low temporal resolution, i.e., large time frames are treated as one and aggregated to a single annotation (e.g., 10-minutes time windows [19]). Therefore, short-term deviations in stress levels cannot be modelled with sufficient precision, leading to problems in certain application domains, e.g., scenarios that require *real-time* stress detection.

In contrast to that, the dataset presented in this paper was annotated by experienced psychologists in a time-continuous manner. This allows for the development of stress recognition systems that are more accurate, reactive, and robust than is the case with existing datasets. In addition to that, we analyzed not only self-perceived stress but also the participants' biological stress response in order to validate whether the study setup has indeed been eliciting a stress response. We argue that considering both aspects, namely self-perceived stress and biological stress response, even though being time-consuming, is important in order to provide credible stress labels when employing study setups that have not been already validated. To the best of our knowledge, this two-fold analysis has not been done in any other freely available stress dataset.

Even though multi-modal stress datasets exist, they rarely provide a comprehensive representation of the participants' behavior and lack a multi-modal assessment of physiological stress responses. The majority of the datasets mainly focus on physiological signals, e.g., HRV and EDA. Therefore, they mostly neglect various aspects of non-verbal behavior, like body language [23], [24], [25], that also hold valuable information about perceived stress. In order to reliably assess a person's experienced stress in different environments (e.g. office, home, recreational activities) it is important to acquire a comprehensive representation of the person's response to stress. For instance, focusing exclusively on physiological modalities such as heart rate or electrodermal activity might yield heightened measurements during physical exercise without necessarily indicating a correlate of psychological stress. Similar scenarios can be found when only considering non-verbal behavior. Finally, out of the presented existing stress datasets only two provided multi-modal baseline results for the automatic recognition of stress [19] [13].

Recent research revealed that models for the automatic recognition of stress showed a significant decrease in prediction performance when tested on other datasets that have not been used for training [26]. Therefore, additional datasets are needed that are compatible with already existing ones in terms of available modalities. Having a collection of compatible datasets enables researches to train stress models across multiple datasets resulting in increased generalizability. Therefore, a special emphasis when creating the ForDigitStress dataset has been placed on providing a comprehensive collection of modalities that are compatible with already existing datasets.

The proposed ForDigitStress dataset contains audio, video, skeleton data, facial landmarks including action units as well as physiological information (PPG, EDA). In addition to the raw signals, we also provide already extracted features for HRV and EDA as well as established feature sets like GEMAPS [27] and OpenPose [28]. Furthermore, this dataset contains pupillometry data, which is a mostly

overlooked modality for the recognition of stress. As prior work suggests [29], [30], [31], there are correlations between various affective states and pupil dilation. Also, collecting pupillometry data can be done unobtrusively by using existing eyetrackers or even laptop webcams [8]. Therefore, we believe that incorporating pupillometry data can benefit multiple stress-related use cases where eye-tracking is a reasonable option. The dataset provides already extracted pupil diameter as well as close-up infrared videos of the eye. Based on the close-up videos we trained an autoencoder and extracted the latent space features that represent an abstracted version of the eye. Those features are also made available as part of the dataset.

3 Dataset

3.1 Design Principles

3.1.1 Setting

The main requirement for the study setting has been to elicit stress and emotional arousal in participants. Moreover, the setting should reflect a familiar real-world scenario. Therefore, we opted for a remote job interview scenario, a typical digital stressor. Performing remote job interviews has become a common procedure in many modern working environments. Job interviews are by their nature a complex stressful social scenario where different aspects of human interaction and perception collide. Previous research has shown that psycho-social stress also occurs in mock job interviews [32], [33]. Figure 1 shows a schematic of the employed study setup. To mimic remote job interviews, participant and interviewer were interacting via two laptops while sitting in two separate rooms. The participants were alone in the room the whole time (except during preparation). This means that no stress caused by social evaluation of a present person was generated.

3.1.2 Procedure

The study procedure consisted of two parts. Prior to the day the mock job interview took place the participants sent their curriculum vitae (CV) to the experimenter and filled out an online survey, in which demographic variables and experiences with job interviews were assessed. This survey further included questionnaires where participants' personality traits, coping styles, perceived stress during the last month, as well as trait anxiety and trait-depression were assessed, which are not further considered here, but which are described and evaluated in [34].

On the day of the experiment, participants were invited to the laboratory and were told that physiological reactions during an online job interview would be recorded. Furthermore, they were asked about their dream job and were equipped with PPG and EDA sensors and a wearable eye tracker. Then, they had about 15 minutes to prepare for the interview. During this time, the interviewer also prepared for the interviews and thought about questions related to the applicant's CV and dream job. Afterwards, the participant and interviewer were seated in two separate rooms, and the interview started. They interacted with each other over two connected laptops, similar to an online meeting. The interviewer was instructed to ask critical questions to stress the applicant and to induce similarly negative emotions. The interviews were structured and the same areas were always queried but with a different focus and related to the specific job the participant applied for. The content of the interviews included questions about the strengths and weaknesses of the applicant, dealing with difficult situations on the job, salary expectations, willingness to work overtime, and inconsistencies in the CV. In addition, tasks related to logical thinking were asked as well as questions about basic knowledge in the areas of mathematics and language. A typical interview followed this pattern [34]:

1) *Reception*
2) *Self-introduction of the candidate (e.g. "How would you describe yourself?")*
3) *Interrupting the introduction (e.g. "Don't share information with me that I can also find in your CV.")*
4) *Intrinsic motivation (e.g. "Why do you want to work for this particular company/employer?")*
5) *Reason for job change (e.g. "Why are you seeking a new job?")*
6) *Candidate's expectations (e.g. "What do you expect from this job?")*
7) *Self-promotion candidate (e.g. "What qualifies you for this role?")*
8) *Hypothetical situation (e.g. "How would you react if...?")*
9) *Applicant's vision of the future (e.g. "Where do you see yourself in 5 years?")*
10) *Spontaneous task (e.g. "Do you have a pen at hand right now? Sell me this pen!")*
11) *Questions about basic knowledge (e.g. "Translate the following sentence...")*
12) *Salary and working hours expectations (e.g. "Are you prepared to work overtime?")*
13) *Outfit (e.g. "Why are you wearing this outfit today?")*
14) *End*

The guidelines for the interviews are freely available in the Open Science Framework (https://osf.io/5bdyf/).

After the job interviews, participants were asked about their emotions during the job interview. For this, qualitative, semi-structured interviews were used (see https://osf.io/5bdyf/). After reporting the experienced emotions, participants reported whether they felt stressed at any time during the interviews, and were instructed to describe as precisely as possible in which specific situations during the job interviews they felt stressed. This procedure (rating and assignment to specific situations) was repeated for all of the reported emotional states (i.e., shame, anxiety, pride, anger, annoyed, confused, creative, happy, insecure, nervous, offended, sad, surprised). These qualitative interviews lasted about 10 to 20 minutes, depending on the participant.

In order to assess whether the mock job interview did elicit stress in the participants, we collected self-reports as well as saliva samples to determine cortisol levels. Salivary cortisol levels are a measure of the activity of the hypothalamic-pituitary adrenal (HPA) axis. Increased cortisol levels can be observed when a person is exposed to stress [35], especially in social-evaluative situations, and are a typ-

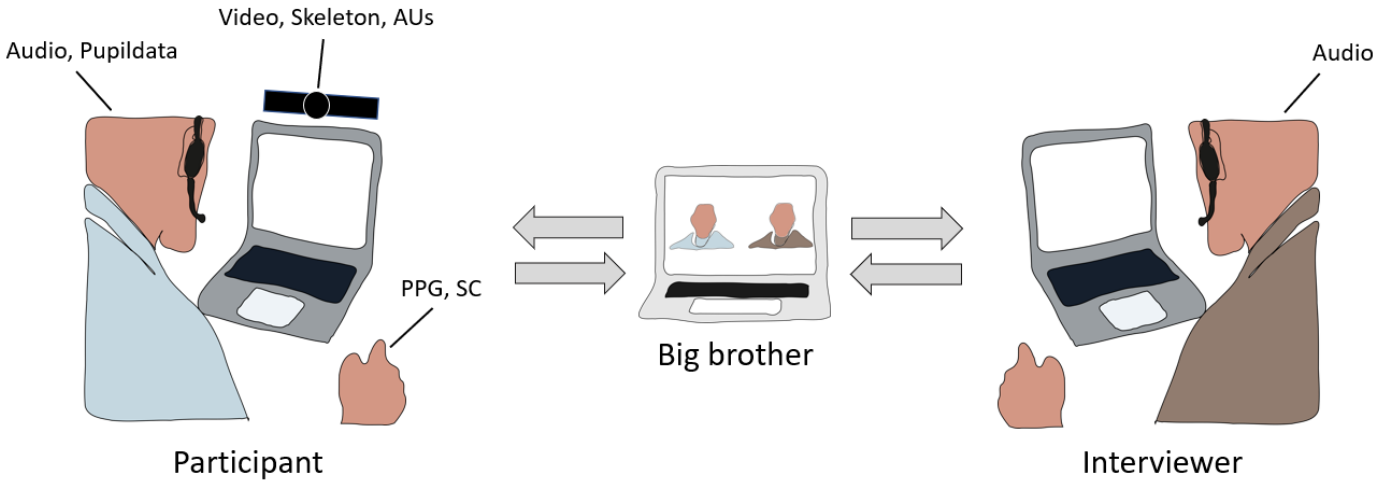


Fig. 1. Overview of the study setup including information on the different modalities that have been recorded during the interviews. Participant and interviewer were seated in different rooms and interacting remotely with each other. A third computer was acting as an observer to unobtrusively monitor the interaction between the participant and interviewer.

ical marker in research on acute biological stress responses (e.g., [36], [37]). Cortisol levels were, therefore, considered an adequate measure to investigate the participant's biological (i.e., HPA axis) response to the digital job interview. After a person has been exposed to a stressor, cortisol levels do not increase instantly. Peak levels are usually found after 10 to 30 minutes [35] after psycho-social stressors of short duration (e.g., the TSST, which is similar to a job interview scenario, because both include strong socially-evaluative components). After this, cortisol levels return to baseline levels. The samples of participants that have been stressed by the job interview - in the sense of an activation of the HPA axis - will show an increase in cortisol levels until they reach a peak followed by a decrease back to their baseline levels. Therefore, salivary cortisol was assessed as a measure of biological stress. For saliva collection, salivettes (Sarstedt, Numbrecht) were used. Each participant provided six saliva samples at different time points. Figure 2 displays an overview of the timing of saliva sample collection during the study. The first sample was collected at the beginning of the study and the second at the end of the preparation phase (i.e., immediately before the actual job interview started). Those two samples were separated by about 15 minutes in order to assess the baseline cortisol level before the participant was exposed to the stressor, i.e., the job interview. The next four samples were collected immediately after the job interview, 5 minutes, 20 minutes, and 35 minutes after it to cover the cortisol increase, its expected peak, and its return to baseline. During each saliva sampling, participants rated their current stress level on a 10-point Likert scale with the anchors "not stressed at all" and "totally stressed", which have been used in previous studies ([38]).

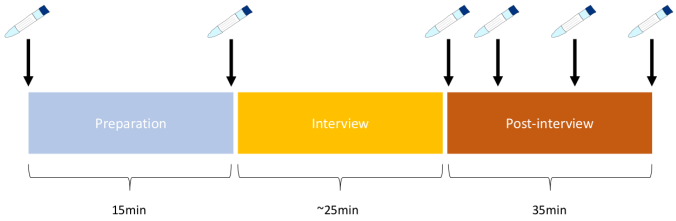


Fig. 2. Overview of the timing of saliva sample collection during the different stages of the study.

3.2 Recording System

Various sensors were used to record the participants' physiological responses. For recording and streaming the participant's data, we employed a Microsoft Kinect 2. The Microsoft Kinect 2 supports Full HD video captures as well as optical motion capturing to extract skeleton and facial data. Moreover, the built-in microphone was used to record ambient sound data. In addition to that, the participants were equipped with an ordinary business USB headset from Trust. Furthermore, the IOM-biofeedback sensor was used to collect PPG and EDA data. Finally, participants were wearing a Pupil Labs eyetracker to record closeup videos of their eyes. All sensors were connected to a Lenovo Thinkpad P15. The setup for the interviewers only consisted of audio recorded with the same Trust USB Headset and video from the built-in Lenovo Thinkpad P15 webcam. A schematic overview of the recording setup is displayed in Figure 1. The participant and interviewer were seated in two different rooms and were interacting remotely with each other through the two laptops. In a third room, another computer was set up to act as an observer. This way the interaction between the participant and the interviewer could be monitored unobtrusively. In order to keep the recorded signals in synchrony we implemented a Social Signal Interpretation Framework (SSI) [39] pipeline. SSI includes an interface for the development of online recognition systems from various sensory devices.

3.3 Collected Data

Data of N = 40 healthy participants (57.5% female, 40% male, 2.5% diverse) was included in the data set. Mean age was 22.7 ± 3.2 years (min: 18, max: 31). Mean body-mass-index (BMI) was 23.2 ± 4.1 kg/m^2 (min: 17.9, max: 37.7; 1

Sensor	Filename	Signal
IOM	bvp.csv	PPG
	sc.csv	EDA
Kinect	video.mp4	HD Video
	skel.csv	Skeleton Data
	face.csv	Facial Points
	head.csv	Head Position
	au.csv	Action Units
	kinect.wav	Audio (room)
Headset	close.wav	Audio (close-talk)
Eyetracker	eye.mp4	Video (close up eye)

TABLE 2
List of recorded files available for download

missing). In total 56 hours and 24 minutes of multi-modal data have been recorded. An overview of all the recorded files is displayed in Table 2.

3.4 Annotation

Two experienced psychologists annotated the recorded sessions frame by frame based on the participants' reports and the content of the interviews. Categories for the annotations were the categories from the questionnaire, i.e. stress as well as the reported emotions like shame, anxiety, anger, and pride. In total, 21 hours and 26 minutes of data were annotated. Figure 3 displays the overall label distribution for the occurred emotions. In the first step, the two psychologists independently annotated the 40 videos with the NOVA tool [40]. During the annotation process, the job interview videos were examined regarding stress and different emotions. The first round of annotation was created based on the observable verbal and non-verbal behavior of the participants. Each video was carefully watched and as soon as an emotion-specific behavior appeared or the participant's behavior indicated stress, an annotation was created for the corresponding emotion or stress. For example, stress was concluded if a person was sitting very restless in their seat. In the second step, the annotations were supplemented with information from the self-reports of the interviewees. For every visible or reported feeling of stress, a discrete label was created for the corresponding frames. Emotions were annotated accordingly. For example, if a participant reported the emotion of shame in a certain situation during the interview, an annotation for the emotion of shame was created for that sequence. There were no disagreements between the psychologists' ratings and the participants' self-reports, i.e., for every situation that was assigned to stress or an emotion by the participants, a time window could be assigned by the psychologists and a corresponding annotation could be created. In the last step, disagreements in the annotations were discussed by the two psychologists. These only affected the annotations, which were created based on observable verbal and non-verbal behavior, as there has been no disagreement for the self-reported emotions and stress. After a detailed discussion and in the case of an agreement between the two psychologists regarding the person's stress, both annotations were adjusted to ensure a standardized representation. The same procedure was also used for the annotations of the emotions. If both psychologists could clearly identify the same emotion after the discussion, the annotations were adjusted accordingly. But if, after the discussion, there were still differences in the perception of the presence of an emotion or the extent of stress, the original annotations of the two psychologists were maintained for the corresponding situations without making any further changes. The annotations were conducted in multiple sessions. On average, each session lasted two hours, and if signs of distraction were noticed, a break was taken, or the annotation continued on the following day. The videos were viewed multiple times by the two psychologists with intervals in between.

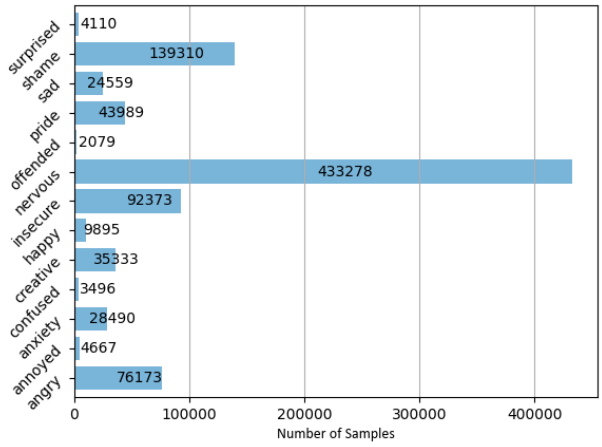


Fig. 3. Number of samples per occurred emotion.

A screenshot of a loaded recording session from the dataset is shown in Figure 5. The screenshot displays a situation during the interview phase where the participant experienced stress. As a consequence, changes in the physiological signals as well as nonverbal behavior could be observed, e.g. pupil dilation, activation of specific action units (in this case the lip corner puller), changes in heart rate variability, and electrodermal activity.

In order to measure the quality and reliability of the value-discrete and time-continuous annotations, we calculated the interrater agreement between the two psychologists using Cohen's Kappa (see Figure 4). The majority of the annotations have shown a strong to almost perfect agreement following the interpretation for Cohen's Kappa.

3.5 Feature Extraction

The recorded raw data has been used to extract features that are valuable for stress recognition. The following section gives an overview of the extracted features as well as additional information regarding the extraction process. Moreover, the presented features are also available for download.

3.5.1 EDA

Features derived from skin conductance (SC) as a measure for EDA are widely used for stress recognition [15], [17], [19], [41]. The EDA signal can be decomposed to skin conductance level (SCL) and skin conductance response (SCR) [19], [42]. SCL or the tonic component is the slow-changing part of the EDA signal. SCR or the phasic component are the rapid changes as a response to a specific stimulus. First, we remove the high-frequency noise by applying

TABLE 3
List of features extracted from various modalities

Modality	Features	Description
Action Units	Jaw	Intensities of action units JawDrop (AU26), JawSlide (AD30)
	Lips	Intensities of action units LipPucker (AU18), LipStretcher Right/Left (AU20), LipCornerPuller Right/Left (AU12), LipCornerDepressor Right/Left (AU15), LowerLipDepressor Right/Left AU(16),
	Cheeks	Intensities of action units CheekPuff Right/Left (AU13)
	Eyes	Intensities of action units EyeClosed Right/Left (AU43), EyebrowLowerer Right/left (AU4)
EDA	MeanEDA, StdEDA, MinEDA, MaxEDA, RangeEDA	Mean, Standard deviation, Min, Max, Dynamic Range of EDA signal
	SlopeEDA, MeanDeriv, StdDeriv	Slope of EDA signal, Mean and Standard deviation of 1st derivative of EDA signal
	MeanSCR, StdSCR, MeanSCL, StdSCL	Mean and Standard deviation of SCR and SCL components
	CorrSCL	Correlation of SCL with time
	PeaksSCR, AmplitudeSCR, DurationSCR, AreaSCR	Number of peaks, Sum of peak amplitudes, sum of peak durations and sum of area under the peaks of the SCR signal
PPG	HR	Number of peaks in 1 minute
	MeanNN, MedianNN, MadNN	Mean, Median, Median absolute deviation of HRV
	StdNN, CVNN, IQRNN	Standard deviation, Coefficient of Variation, Inter-Quartile Range of HRV
	RMSSD, StdSD	Root Mean Square and Standard deviation of successive differences of P-P intervals
	pNN50, pNN20	Percentage of successive differences of P-P intervals $> 50\ ms$ and $> 20\ ms$
	TINN, HTI	Triangular Interpolation of HRV histogram, HRV Triangular Index
	LF, HF, LF/HF	Low Frequency ($0.04\ Hz - 0.15\ Hz$) and High Frequency ($0.15\ Hz - 0.4\ Hz$) power
	LFn, HFn	Normalized low and high-frequency power, LF/total power, HF/total power
	SD1, SD2, SD1/SD2	Spread of HRV points on Poincaré plot along identity line and perpendicular to it
	S	Area of the ellipse formed by HRV points in the Poincaré plot
Audio	GEMAPS	Pitch, Jitter, Formant 1-3 frequency and relative energy, Formant 1 bandwidth, Shimmer, Loudness, Harmonics-to-noise ratio(HNR), Alpha Ratio, Hammarberg Index, Spectral Slope, Harmonic difference H1-H2, Harmonic difference H1-A3
Body Keypoints	Kinect	x, y, z position and rotation of head, forehead, nose, left/right ear, chin, neck, torso waist, left/right shoulder, left/right elbow, left/right wrist, left/right hand, hip left/right, left/right knee, left/right ankle, left/right foot
	OpenPose	x, y position of the nose, left/right eye, left/right ear, neck, left/right shoulder, left/right elbow, left/right wrist, hip left/middle/right, left/right knee, left/right ankle, left/right big toe, left/right small toe, left/right heel
Eye	Pupil features	pupil diameter, latent space features extracted from an autoencoder trained on the close-up videos of the eye

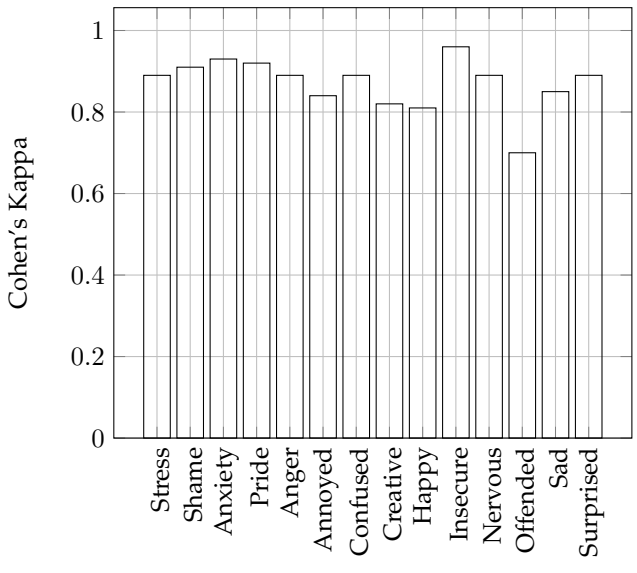


Fig. 4. Average Cohen's Kappa calculated for stress and each emotion to map the interrater agreement between the two psychologists.

a 5 Hz low-pass filter [19], [42]. We use the filtered signal to calculate statistical features [19], [41], [43] like mean, standard deviation, dynamic range, etc.. We compute the SCL and SCR components using the cvxEDA decomposition algorithm [44]. In addition to the various statistical features of SCL and SCR signals, we also compute features derived from the peaks in the SCR signal [15]. We compute a total of 17 features (see Table 3) from a 60 seconds long EDA input signal. The 60 seconds time frame was chosen based on similar research that achieved excellent results with it [45].

3.5.2 PPG

As demonstrated in previous studies [19], [46], the PPG signal can be used to derive HRV (Heart Rate Variability) features for predicting stress. We compute 22 PPG-based HRV features which are listed in Table 3. To derive the HRV from PPG, we detect the Systolic Peaks (P) from the input signal. The first step is to remove baseline wander and high-frequency noises from the raw PPG signal. We use a band-pass filter ($0.5 - 8$ Hz) to reduce the noise and enhance the peaks [47]. Next, we use a peak finding algorithm to detect peaks such that (a) their amplitudes

Fig. 5. An instance of a recorded session loaded in NOVA.The top row displays the eyetracking video alongside the video recording of the participant. Below that several feature streams are displayed: HRV feature stream, EDA, action units, and GEMAPS audio features. At the bottom, two discrete annotation tiers are shown. The first tier displays situations where the participant experienced stress while the second tier displays the interview phase.

are above a specified threshold and, (b) consecutive peaks are sufficiently apart. The amplitude threshold is set to the mean of the 75 percentile and 90 percentile of the peak heights in the input signal. In a previous exercise stress study [48] involving healthy participants of varying ages, the maximum heart rate recorded was 3 beats per second (180 beats per minute). Hence, we set the minimum time between two consecutive peaks as 0.333 seconds. We use 60-second long PPG segments to detect the peaks and compute the HRV signal. We compute various HRV features [19], [26], [41], [43], [49], [50] from the time domain, frequency domain, and Poincaré plots.

3.5.3 Body keypoints

Prior studies have established the value of body language and body behaviour for the recognition of stress [23] [24] [25]. Therefore, our study setup included a Microsoft Kinect2 to extract 3D body data. This data provides information about 25 joints, consisting of position in 3D space, orientation of the joints in 3D space as well as a confidence rating in regard to the tracking performance. Even though the Microsoft Kinect2 has been used in prior studies in the context of stress recognition [23], [24], [25] we aimed to provide additional body data in order to enable others to utilize the provided dataset across multiple datasets. Therefore, we extracted the OpenPose [28] features from the recorded HD video displaying the participant. OpenPose is a widely used state-of-the-art framework for the detection of human body key points in single images. It is important to point out that OpenPose solely returns the body key points in 2D space, therefore, losing some information when compared to the Microsoft Kinect2 data. However, in order to extract the OpenPose features no special hardware is required and the data of a simple camera is sufficient. Also, due to the study setup, not all joints could be successfully tracked, as the participants were sitting and their lower body was concealed by the table. Therefore, only the features corresponding to the upper body joints provide reliable information.

3.5.4 Action units

Facial expressions play an important role in communicating emotions and therefore are frequently used for the automatic detection of affective states [51] [52]. Furthermore, recent studies have utilized facial action units to successfully predict human stress [24] [53] [54]. We extracted 17 facial action units (see Table 3) provided by the Microsoft Kinect2. In addition to that, we also extracted the OpenFace2 [55] features that consist of facial landmarks, head pose, facial action units, and eye-gaze information. Similar to OpenPose, those features can be extracted from any video data.

3.5.5 Audio features

Knapp et al. [56] argue that emotions are reliably transported by the voice. Indeed it is a well-established fact that the acoustic characteristics of speech e.g. pitch and speaking rate are altered by emotions [57]. Moreover, vocal signs of stress, such as an increase in fundamental frequency [58] or changes in vocal tremor [59], are mainly induced by negative emotions [60]. An increase in fundamental frequency during an episode of experienced stress can also be seen in Figure 5 (GEMAPS feature stream; 4th row). Multiple studies were able to show that it is possible to automatically detect stress with acoustic features [61] [60] [62] [63]. In order to provide meaningful acoustic features

we chose to extract the GEMAPS features [27]. One of the main objectives of the GEMAPS feature set has been to provide access to a comprehensive and standardized acoustic feature set. It contains frequency and energy-related features like pitch, jitter, shimmer and loudness, as well as spectral features, e.g., Hammarberg Index and harmonic differences. We calculated the features over a one-second time window.

3.5.6 Pupil features

Responses of the pupil, like pupil dilation, are closely related to subjective and physiological stress responses (e.g., the activation of the hypothalamic-pituitary adrenal (HPA) axis; [64] [65]). Furthermore, a recent study has shown that pupillometry is a suitable tool to measure arousal during emotion regulation after an acute stressor [66] [65]. However, this modality has not yet been paid much attention in established affective computing datasets [67]. Therefore, part of our study setup has been a wearable eye tracker that provides close-up video data of the participant's eye. From those videos, we automatically extracted the pupil diameter by employing the extraction pipeline described in [8]. In addition to that, we also trained an autoencoder on the close-up eye videos in order to extract the corresponding latent space features. The latent space features contain an abstract representation of the eye. Figure 6 displays the original input image of the eye and the reconstructed output image produced by the autoencoder below. During the encoding and decoding process, barely any loss of information occurred as the input image and corresponding output image are almost identical. This is a strong indicator that the autoencoder has learnt meaningful features to accurately translate the image into and out of the latent space. The resulting feature set consists of 512 parameters corresponding to the size of the latent space.

Fig. 6. Examples of reconstructed images. The top row displays the original input image, while the bottom row shows the images reconstructed by the autoencoder.

3.6 Availability

The ForDigitStress dataset is freely available for research and non-commercial use. Access to the dataset can be requested at https://hcai.eu/fordigitstress. The dataset is organized in sessions with a total size of approximately 360 GB.

3.7 Automatic Stress Detection

3.7.1 Dimensionality Reduction

As seen from Table 3, numerous features have been extracted from each modality. The size of the input dimension can be a concern for some machine learning techniques, especially when we consider multi-modal input. Therefore, we use PCA (Principal Component Analysis) as it has been shown to reduce dimensionality without a drop in classification performance of machine learning models [68]. We apply PCA for stress models involving individual modalities as well as multi-modal stress recognition models. The length of the feature vectors of action units, EDA, HRV, OpenPose, and GEMAPS was $17, 17, 22, 24, 58$. We retain 95% of the components using PCA, reducing the length of the feature vectors to $10, 9, 10, 8, 19$, respectively. The following approach is applied for combining features for multi-modal stress recognition. We first apply PCA to individual modality features and then combine them (i.e., concatenated the reduced features). The final length of the feature vector is 56 (sum of the length of feature vectors of each modality). Similar to Reddy et al. [68], we perform MinMax normalization between 0 and 1 before applying PCA.

3.7.2 Classifiers

Previous works [19], [43], [69], [70] have demonstrated that many machine learning classifiers such as SVM (Support Vector Machine), KNN (K-Nearest Neighbors) and RFC (Random Forest Classifier) can achieve good stress recognition performance. Recent works [26], [71] have shown that neural networks perform better than popular machine learning classifiers in feature-based stress recognition. We train the following classifiers as a baseline for our dataset.

- **KNN** This machine-learning technique classifies samples based on the labels of the nearest neighbouring samples. The neighbouring samples are determined using the Euclidean distance between them. We use $K = 50$ neighbouring samples to classify the samples. This parameter was chosen by extrapolating the threshold value for stable performance, $K = 10$, for WESAD dataset [72]. Considering the stress duration and number of participants, our dataset contains almost 5 times more data than WESAD, which is reflected in the chosen K parameter.
- **Feed-forward Neural Network** This is a Multi-Layer Perceptron with an input layer, two hidden layers, and a prediction layer. Since the size of the input varies depending on the modalities, we have a varying number of nodes in the hidden layers. We set the number of nodes in the first hidden layer as half of the input size, rounded up to a multiple of 2. The number of nodes in the second hidden layer is half of the first layer. The activation function for hidden layers is ReLU (rectified linear unit). The prediction layer has a single node with Sigmoid activation to discern between stress and no-stress classes. We avoid over-fitting by using a dropout layer (dropout rate $= 0.2$) after the input layer.
- **RFC** This is an example of an ensemble classifier that trains a number of decision tree classifiers on subsets

of the training set. This training technique controls over-fitting. Hence, the RFC achieves better overall performance, even if the individual decision trees are weak. In our evaluations, we use an RFC with 100 decision trees (or estimators) and 50 minimum samples for splitting a node.

- **SVM** This is a popular supervised learning technique that often achieves good stress recognition performance. Similar to previous works [43], [70], we use the Radial basis function (Rbf) as the kernel function for our SVMs.
- **LSTM** In order to incorporate temporal context for the stress classification, we trained an LSTM (Long Short-Term Memory) classifier. An LSTM is a type of recurrent neural network that is usually used to process sequential data and is able to capture temporal relationships. The LSTM model consists of one LSTM layer with 256 units and a time step size of 50 samples followed by one hidden layer with the number of nodes as half of the input size, rounded up to a multiple of 2. The activation for the LSTM layer is TanH while the activation function for the hidden layer is ReLU. The prediction layer has a single node with a sigmoid activation function to discern between stress and no-stress classes. We avoid over-fitting by using a dropout layer (dropout rate = 0.4) after the LSTM layer. The time step size of 50 samples was identified after an experiment where we incrementally increased the step size starting with 25 samples (i.e., one second of data). A time step size of 50 samples achieved the best results regarding accuracy.

The feed-forward neural networks and LSTMs were implemented using Tensorflow (version 1.15.0). We use the SGD optimizer (learning rate $=$ 0.001) and binary cross-entropy loss. We train them for 100 epochs while employing early stopping with a patience of 15. All other machine-learning models were trained using Scikit-learn (version 1.0.2). We balanced our training set by randomly down-sampling the no-stress class depending on the number of stress samples annotated for each participant. Further, before training the models all feature vectors have been normalized between 0 and 1. The training procedure for the autoencoder features extracted from the eye tracker video data differs from the other modalities. First, some participants manipulated the eye tracker by accidentally bumping into it and changing the alignment of the built-in camera. In some cases, this resulted in uncaptured eyes. In these cases, it was not possible to re-align the eye-tracker so as not to disturb the study procedure. Therefore, the models could only be trained on a subset of the recorded data. The subset contains 19 sessions for training the model. For that reason, we report the results separately from the baseline results. Apart from the reduced training data, the procedure for training these models was similar to the other classifiers.

3.7.3 Evaluation Metrics

Similar to previous work [19], [26], we use accuracy and f1-score as the performance metrics to evaluate our stress models. To assess the generalizability of our models on data from unseen users, we perform LOSO (leave-one-subject-out) evaluations.

4 Results

4.1 Automatic Stress Detection

We evaluate our dataset on a binary stress recognition task (stress vs. no stress). The dataset has a sample rate of 25 Hz. We predicted stress for every sample of the annotated data. Popular machine learning techniques such as RFC, KNN, SVM, Feed-forward Neural Networks, and LSTM are trained on features extracted from facial action units, EDA, HRV, OpenPose, and GEMAPS. The results of our LOSO evaluation are presented in Table 4.

Combining modalities yields better stress recognition performance than individual modalities. The LSTM model achieved the best stress recognition performance ($F1 = 90.2\%$, $Accuracy = 91.7\%$). A previous study [73] reported similar findings, where a CNN-LSTM model outperformed simple machine learning models on a custom stress dataset. Among the simpler models, feed-forward NN achieved better performance ($F1 = 88.1\%$, $Accuracy = 88.3\%$). This result is in line with the observations of related work [26], [71], [74], [75], where a simple feed-forward network achieved better performance than other machine learning models (SVM, RFC, etc.) on multimodal stress datasets (e.g., WESAD, SWELL).

When considering stress recognition using a single modality for the models that are not considering temporal context, HRV features yield the best results across classifiers, followed by facial action units and OpenPose features. The GEMAPS and EDA features rank the lowest in stress recognition performance, achieving $15 - 20\%$ lower f1-score and accuracy. When employing LSTM models which incorporate temporal context the results differ. The best performance for single modality stress recognition was achieved with the GEMAPS features, whereas the worst was scored with the OpenPose features. HRV and action units continued to provide good recognition scores.

As mentioned in subsection 3.7 we also trained classifiers on the extracted eye autoencoder features. Due to the reduced training data we report the results separately. The best performance was achieved by the LSTM model with an f1-score of 68.3 and an accuracy of 70.2% followed by the neural network model with an f1-score of 54.8 and an accuracy of 62.0%. The non deep learning approaches struggled to reach accuracy scores above chance. The KNN classifier achieved an f1-score of 47.2 and an accuracy of 48.9%. The SVM model had an f1-score of 45.3 and an accuracy of 47.9%. The worst performance was from the RFC with an f1-score of 39.1 and an accuracy of 49.5%.

4.2 Biological Stress

As a manipulation check, i.e. to prove whether our job interview scenario indeed induced stress, biological and perceived stress were measured at 6 points in time (2 before and 4 after the job interview). Cortisol levels as a marker for biological stress significantly changed during the whole session (Figure 7A; $F(5, 190) = 3.19$, $p = 0.009$). They were highest 5 minutes after the job interview and then decreased

TABLE 4
Evaluation of classifiers on different modalities for the binary stress recognition task. Presented are mean F1-scores and accuracies. The standard deviations are displayed in brackets.

	RFC		KNN		SVM		Simple NN		LSTM	
Features	F1 (SD)	Acc (SD)	F1 (SD)	Acc (SD)	F1 (SD)	Acc (SD)	F1 (SD)	Acc (SD)	F1 (SD)	Acc (SD)
AU	71.4 (17.5)	73.6 (14.8)	70.7 (16.3)	73.1 (13.8)	75.6 (14.5)	77.2 (12.0)	76.5 (14.7)	78.0 (12.2)	**80.1** (19.8)	**83.1** (15.7)
EDA	54.2 (16.3)	57.1 (14.6)	54.6 (10.4)	55.2 (10.3)	57.6 (20.8)	58.9 (20.3)	60.2 (22.1)	61.3 (21.5)	**72.5** (21.6)	**75.9** (16.9)
HRV	74.5 (18.5)	75.9 (17.2)	72.6 (14.6)	73.2 (14.0)	76.1 (22.1)	77.7 (20.5)	78.4 (21.3)	79.7 (19.9)	**81.0** (22.2)	**82.6** (19.5)
OpenPose	59.4 (19.7)	63.6 (16.3)	67.0 (19.7)	69.5 (17.0)	69.8 (20.4)	73.4 (16.0)	**76.4** (22.3)	**79.5** (17.2)	68.4 (30.0)	73.0 (25.6)
GEMAPS	52.1 (7.0)	55.9 (4.4)	55.1 (7.9)	56.9 (6.5)	57.3 (9.0)	58.9 (7.7)	58.7 (9.1)	60.3 (7.6)	**84.0** (23.9)	**86.7** (23.9)
All	81.3 (15.4)	82.0 (14.2)	74.7 (14.9)	75.5 (13.8)	83.8 (17.4)	84.5 (16.5)	88.1 (14.1)	88.3 (13.4)	**90.2** (18.4)	**91.7** (14.5)

to baseline levels 35 minutes after the stressor. A similar time course was found for perceived stress, which was highest immediately after the job interview and decreased to baseline afterwards (Figure 7B; $F(5, 190) = 39.82$, $p < 0.001$).

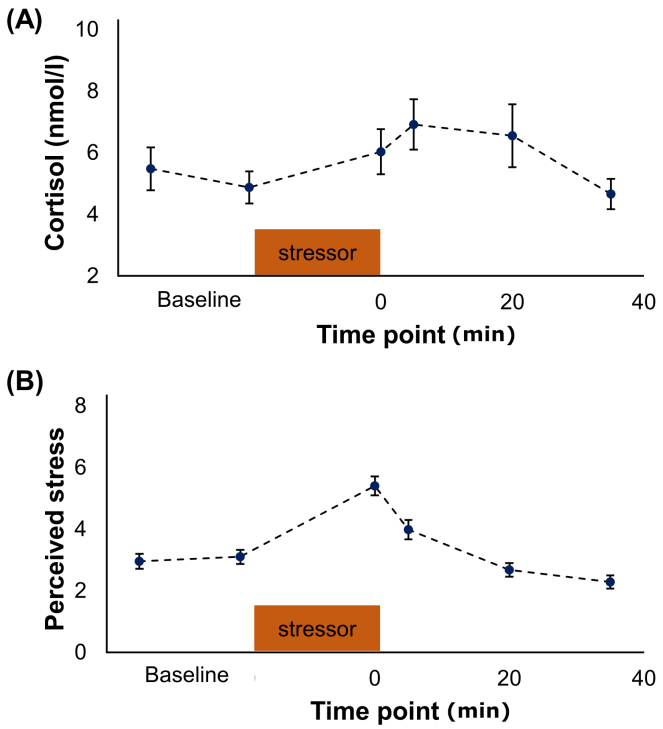


Fig. 7. Time course of cortisol levels (A) and perceived stress (B) during the whole session.

5 Discussion

In order to establish a baseline on our dataset for the automatic recognition of stress we trained several machine learning models on different modalities. Overall, we found that a fusion of the action units, EDA, HRV, OpenPose and GEMAPS features that have been reduced in dimensionality - by employing PCA - achieved the best accuracy and f1-scores with 91.7% and 90.2. Throughout our experiments, the LSTM model outperformed the other classifiers in most of the modalities except for the OpenPose features. Here, the simple NN yielded the best performance. This emphasizes the importance of temporal context for the automatic detection of stress. In single-modality stress recognition, the HRV features achieved good results across all classifiers. In fact, for the models that don't consider temporal context, the HRV features achieved the best results among the different single modalities. This is in line with existing research that identified HR and HRV as excellent measures for predicting stress [76], [77]. Moreover, models trained with action units achieved similar results, i.e., 83.1% compared to 82.6% for the HRV features. Another well-established modality to detect stress is EDA [76]. Models solely trained on EDA features were able to achieve accuracy scores of up to 94.62 % in a 2-class and up to 75 % on a 4-class pattern recognition task on a modified Trier Social Stress dataset [78]. Interestingly, in our experiments, the models trained on the EDA features were the ones having the second-worst accuracies and f1-scores. One reason for that observation could be that existing datasets often aggregate larger time frames to one label whereas we worked with time-continuous annotations with a high temporal resolution. This could be a problem when working with EDA as there is a delay between the sympathetic nervous systems stimulation and the corresponding EDA response [79]. Therefore, the EDA features could still represent a non-stressed state due to the delay for situations identified as stress. This could potentially be mitigated by either shifting the signal corresponding to the delay or calculating the EDA features over a longer time window. Further investigations should be conducted in future work to check whether following those approaches leads to better classification performance. Finally, when training the classifiers on the GEMAPS features, the consideration of temporal context had the biggest influence. While for the models that do not incorporate temporal context, the GEMAPS features yielded the worst results, the feature set achieved the overall best single modality scores when training an LSTM. The LSTM model trained on GEMAPS achieved an f1-score of 84.0 and an accuracy of 86.7%.

In general, our experiments showed excellent f1-scores and accuracies for the automatic recognition of stress but also revealed quite high standard deviations. This means that the performance for the single splits substantially differed. After manual inspection of the f1-scores and accuracies, we found that only a small subset of the sessions showed substantially worse results. This was predominantly due to imprecisions in the feature extraction process, e.g., the OpenPose feature extraction partially resulted in misaligned head tracking.

In addition to the baseline models, we trained the different classifiers on a reduced dataset containing the extracted eye features including pupillometry features. In

those experiments the KNN, RFC, SVM were not able to achieve results above chance. The simple neural network achieved an accuracy of 62.0%. The best performance was achieved by the LSTM model with an accuracy of 70.2%. This discrepancy between deep learning models and conventional models is most likely due to the complexity of the autoencoder features. The autoencoder features set consist out of 512 features which is almost 10 times more than the GEMAPS features which is the second largest feature set employed. Further, the improvement when incorporating temporal context by employing a LSTM model indicates that snapshots of changes in pupil diameter and eye movement, while providing some information about experienced stress, are not sufficient enough to automatically predict stress. Still, the results show that features extracted from close-up eye video data hold relevant information for the recognition of stress. Considering that there is only very limited research available [8] [14] that used close-up eye features, including pupillometry features, to automatically detect stress, this experiment highlights the usefulness of such features. Features derived from the movement of the eye as well as changes in pupil size are a promising, non-invasive modality for the automatic recognition of stress.

In order to validate whether our digital job interviews are a suitable scenario for inducing stress, we measured biological as well as perceived stress during the study. Salivary cortisol levels were used as a marker for biological stress. We found a significant change in cortisol levels and perceived stress throughout the study. Peak cortisol levels were observed 5 minutes after the interview whereas perceived stress was found to be highest immediately after the interview. The delay of peak cortisol levels in comparison to perceived stress ratings is due to the fact that it takes some time for the body to release cortisol. In order to reach peak cortisol levels it usually takes 10 to 30 minutes [35]. This delay can be observed in Figure 7. Overall, the results show that mock digital job interviews are a reliable scenario to induce stress (biological and perceived) in participants. Furthermore, it was found that female participants experienced the scenario as more stressful than male participants. Cortisol peaks were higher for participants who experienced the situation as a threat in comparison to participants who experienced it as a challenge (see [34] for further details).

Our dataset includes many of the stress response modalities that are widely used in other stress datasets. So, our dataset holds promise in the development of more robust and high-performing stress detection models, especially through merging of datasets. For example, a recent work [75] showed that the HRV models trained on our dataset perform equally well on other social stress datasets (WESAD and VerBIO). Notably, these datasets differed from our dataset in many factors such as stress intensity, elicitation method, and sensor brands. Moreover, merging the data from the three datasets resulted in an improved stress detection performance. Their findings show the compatibility of our dataset with existing social stress datasets, thus highlighting the potential of our dataset towards developing a generalizable stress detection model.

6 Conclusion

In this paper, we present a comprehensive multi-modal stress dataset that employs a digital job interview scenario for stress induction. The dataset provides signals from various sources including audio, video, body skeleton, facial landmarks, action units, eye tracking, physiological information (PPG, EDA), as well as already extracted features like GEMAPS, OpenPose, pupil dilation, and HRV. In total, 40 participants have been recorded, resulting in approximately 56 hours of multi-modal data. Moreover, the dataset contains discrete annotations created by two experienced psychologists for stress and emotions that occurred during the interviews. The inter-rater reliability for the individual stress and emotion labels showed a substantial to almost perfect agreement (Cohen's $\kappa > 0.7$ for all labels). Based on the stress annotations, several machine learning models (SVM, KNN, RFC, NN, LSTM) were trained to predict stress vs. no-stress. The best single modality performance of 86.7% was achieved by an LSTM trained on the GEMAPS features. The best stress recognition performance ($F1 = 90.2\%$, $Accuracy = 91.7\%$) was obtained by training an LSTM on all modalities.

Moreover, we validated whether the digital mock job interviews are capable of inducing stress by assessing salivary cortisol levels and perceived stress. The analysis revealed a significant change in cortisol levels and perceived stress throughout the study. Therefore, we conclude that digital mock job interviews are well-suited to induce biological and perceived stress.

In summary, the dataset presented in this work provides the research community with a comprehensive basis for further experiments, studies, and analyses on human stress. Due to the multi-modality of our dataset, we provide the possibility for cross-corpus validation for a multitude of existing stress datasets. Therefore, this dataset contributes to the overall goal of building more robust and generalizable stress recognition models.

In future work, we plan to establish an additional baseline for the automatic detection of emotions that occurred during the interviews. For this purpose, we plan to extend the dataset by continuous valence and arousal annotations. Moreover, we aim to investigate the potential of transformer architectures that have shown promise for assessing valence and arousal in emotion recognition tasks [67].

The dataset presents valuable opportunities for advancing the understanding of stress detection in specific contexts, particularly during job interviews. By analyzing the types of questions posed by the interviewer, researchers can refine stress prediction models, enhancing their accuracy [80].

Moreover, integrating stress recgonition models into job interview training with virtual characters [81], [82] could significantly benefit users. Training systems extended by the models offer a realistic simulation of high-pressure situations, enabling individuals to develop and practice stress management strategies. Such preparation could be instrumental in helping them project greater confidence in actual job interviews, thereby potentially improving their performance in these critical assessments.

7 Ethics

The study has been approved by the local Ethics Committee of the FAU (protocol no.: 21-408-S). All participants gave written and informed consent for participation and for publication of their data. Moreover, the presented study has been approved by the data protection officer of the University of Augsburg.

Acknowledgements

This work presents and discusses results in the context of the research project ForDigitHealth. The project is part of the Bavarian Research Association on Healthy Use of Digital Technologies and Media (ForDigitHealth), which is funded by the Bavarian Ministry of Science and Arts. Linda Becker was funded by the Emerging Talents Initiative of the Friedrich-Alexander-Universität Erlangen-Nürnberg. We thank Leonie Bast, Steffen Franke, and Katharina Hahn for data collection.

References

[1] A. P. Association, "Stress," https://www.apa.org/topics/stress, accessed: 2023-01-13.

[2] H. Yaribeygi, Y. Panahi, H. Sahraei, T. P. Johnston, and A. Sahebkar, "The impact of stress on body function: A review," *EXCLI journal*, vol. 16, p. 1057, 2017.

[3] L. Becker and N. Rohleder, "Time course of the physiological stress response to an acute stressor and its associations with the primacy and recency effect of the serial position curve," *PLoS One*, vol. 14, no. 5, p. e0213883, 2019.

[4] A. P. Association, "Stress effects on the body," https://www.apa.org/topics/stress/body, accessed: 2023-01-13.

[5] H. C. Kaltenegger, L. Becker, N. Rohleder, D. Nowak, and M. Weigl, "Associations of working conditions and chronic low-grade inflammation among employees: a systematic review and meta-analysis," *Scandinavian journal of work, environment & health*, vol. 47, no. 8, p. 565, 2021.

[6] H. C. Kaltenegger, L. Becker, N. Rohleder, D. Nowak, C. Quartucci, and M. Weigl, "Associations of technostressors at work with burnout symptoms and chronic low-grade inflammation: a cross-sectional analysis in hospital employees," *International Archives of Occupational and Environmental Health*, pp. 1–18, 2023.

[7] H. C. Kaltenegger, M. Weigl, L. Becker, N. Rohleder, D. Nowak, and C. Quartucci, "Psychosocial working conditions and chronic low-grade inflammation in geriatric care professionals: A cross-sectional study," *Plos one*, vol. 17, no. 9, p. e0274202, 2022.

[8] A. Heimerl, L. Becker, D. Schiller, T. Baur, F. Wildgrube, N. Rohleder, and E. Andre, "We've never been eye to eye: A pupillometry pipeline for the detection of stress and negative affect in remote working scenarios," in *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 2022, pp. 486–493.

[9] T. Baur, I. Damian, P. Gebhard, K. Porayska-Pomsta, and E. Andre, "A job interview simulation: Social cue-based interaction with a virtual character," 09 2013.

[10] A. Heimerl, S. Mertes, T. Schneeberger, T. Baur, A. Liu, L. Becker, N. Rohleder, P. Gebhard, and E. André, "Generating personalized behavioral feedback for a virtual job interview training system through adversarial learning," in *Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I*. Springer, 2022, pp. 679–684.

[11] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition," in *2019 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE, 2019, pp. 1–4.

[12] W. Chen, S. Zheng, and X. Sun, "Introducing mdpsd, a multimodal dataset for psychological stress detection," in *CCF Conference on Big Data*. Springer, 2020, pp. 59–82.

[13] M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "Muse: a multimodal dataset of stressed emotion," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 1499–1510.

[14] Y. Nakashima, J. Kim, S. Flutura, A. Seiderer, and E. Andre, "Stress recognition in daily work," 09 2015.

[15] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[16] E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes *et al.*, "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[17] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 291–298.

[18] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, 2021.

[19] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[20] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "Trier social stress test," *Neuropsychobiology*, 2010.

[21] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of experimental psychology*, vol. 18, no. 6, p. 643, 1935.

[22] E. Turcan and K. McKeown, "Dreaddit: A reddit dataset for stress analysis in social media," *arXiv preprint arXiv:1911.00133*, 2019.

[23] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva, "Using activity-related behavioural features towards more effective automatic stress detection," *PLOS ONE*, vol. 7, pp. 1–16, 09 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0043571

[24] J. Aigrain, S. Dubuisson, M. Detyniecki, and M. Chetouani, "Person-specific behavioural features for automatic stress detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 03, 2015, pp. 1–6.

[25] H. Chen, X. Liu, X. Li, H. Shi, and G. Zhao, "Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–8.

[26] P. Prajod and E. André, "On the generalizability of ecg-based stress detection models," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022, pp. 549–554.

[27] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[28] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[29] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 185–198, 2003.

[30] S. Sirois and J. Brisson, "Pupillometry," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 6, pp. 679–692, 2014.

[31] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[32] J. Campisi, Y. Bravo, J. Cole, and K. Gobeil, "Acute psychosocial stress differentially influences salivary endocrine and immune measures in undergraduate students," *Physiology & Behavior*, vol. 107, no. 3, pp. 317–321, 2012.

[33] P. Gebhard, T. Baur, I. Damian, G. Mehlmann, J. Wagner, and E. André, "Exploring interaction strategies for virtual characters to induce stress in simulated job interviews," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 661–668.

[34] L. Becker, A. Heimerl, and E. André, "Fordigitstress: presentation and evaluation of a new laboratory stressor using a digital job interview-scenario," *Frontiers in Psychology*, vol. 14, p. 1182959, 2023.

[35] D. Bozovic, M. Racic, and N. Ivkovic, "Salivary cortisol levels as a biological marker of stress reaction," *Med Arch*, vol. 67, no. 5, pp. 374–377, 2013.

[36] L. Becker and N. Rohleder, "Associations between attention and implicit associative learning in healthy adults: the role of cortisol and salivary alpha-amylase responses to an acute stressor," *Brain Sciences*, vol. 10, no. 8, p. 544, 2020.

[37] L. Becker, U. Schade, and N. Rohleder, "Evaluation of the socially evaluated cold-pressor group test (secpt-g) in the general population," *PeerJ*, vol. 7, p. e7521, 2019.

[38] L. Becker, H. C. Kaltenegger, D. Nowak, M. Weigl, and N. Rohleder, "Physiological stress in response to multitasking and work interruptions: Study protocol," *Plos one*, vol. 17, no. 2, p. e0263785, 2022.

[39] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 831–834.

[40] A. Heimerl, T. Baur, F. Lingenfelser, J. Wagner, and E. André, "Nova - a tool for explainable cooperative machine learning," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 109–115.

[41] K. Nkurikiyeyezu, A. Yokokubo, and G. Lopez, "Importance of individual differences in physiological-based stress recognition models," in *2019 15th International Conference on Intelligent Environments (IE)*. IEEE, 2019, pp. 37–43.

[42] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2009.

[43] S. Sriramprakash, V. D. Prasanna, and O. R. Murthy, "Stress detection in working people," *Procedia computer science*, vol. 115, pp. 359–366, 2017.

[44] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[45] E. Hosseini, R. Fang, R. Zhang, A. Parenteau, S. Hang, S. Rafatirad, C. Hostinar, M. Orooji, and H. Homayoun, "A low cost eda-based stress detection using machine learning," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 2619–2623.

[46] Y. Cho, S. J. Julier, and N. Bianchi-Berthouze, "Instant automated inference of perceived mental stress through smartphone ppg and thermal imaging," *arXiv preprint arXiv:1901.00449*, 2018.

[47] M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions," *PloS one*, vol. 8, no. 10, p. e76585, 2013.

[48] J. B. Kostis, A. Moreyra, M. Amendo, J. Di Pietro, N. Cosgrove, and P. Kuo, "The effect of age on heart rate in subjects free of heart disease. studies by ambulatory electrocardiography and maximal exercise stress test." *Circulation*, vol. 65, no. 1, pp. 141–145, 1982.

[49] T. Pham, Z. J. Lau, S. Chen, and D. Makowski, "Heart rate variability in psychology: a review of hrv indices and an analysis tutorial," *Sensors*, vol. 21, no. 12, p. 3998, 2021.

[50] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[51] J. F. Cohn and F. De la Torre, "Automated face analysis for affective computing." 2015.

[52] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917305264

[53] G. Giannakakis, M. R. Koujan, A. Roussos, and K. Marias, "Automatic stress detection evaluating models of facial action units," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 728–733.

[54] ——, "Automatic stress analysis from facial videos based on deep facial action units recognition," *Pattern Analysis and Applications*, vol. 25, no. 3, pp. 521–535, 2022.

[55] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018.

[56] M. Knapp, L. and J. Hall, A., *Nonverbal Communication in Human Interaction*. Harcourt Brace, 1997.

[57] J. Tao and T. Tan, "Affective computing: A review," in *Affective Computing and Intelligent Interaction*, J. Tao, T. Tan, and R. W. Picard, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 981–995.

[58] C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter, "Vocal indices of stress: A review," *Journal of Voice*, vol. 27, no. 3, pp. 390.e21–390.e29, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199712002354

[59] E. Mendoza and G. Carballo, "Vocal tremor and psychological stress," *Journal of Voice*, vol. 13, no. 1, pp. 105–112, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199799800642

[60] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "Recognizing stress using semantics and modulation of speech and gestures," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 162–175, 2016.

[61] H. lu, D. Frauendorfer, M. Rabbi, M. Mast, G. Chittaranjan, A. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," 09 2012, pp. 351–360.

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

[63] H. Han, K. Byun, and H.-G. Kang, "A deep learning-based stress detection algorithm with speech signal," in *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, ser. AVSU'18. New York, NY, USA: Association for Computing Machinery, 2018, p. 11–15. [Online]. Available: https://doi.org/10.1145/3264869.3264875

[64] A. A. Zekveld, J. A. van Scheepen, N. J. Versfeld, E. C. Veerman, and S. E. Kramer, "Please try harder! the influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels," *Hearing research*, vol. 381, p. 107768, 2019.

[65] M. Pedrotti, M. A. Mirzaei, A. Tedesco, J.-R. Chardonnet, F. Mérienne, S. Benedetto, and T. Baccino, "Automatic stress classification with pupil diameter analysis," *International Journal of Human-Computer Interaction*, vol. 30, no. 3, pp. 220–236, 2014.

[66] K. Langer, O. T. Wolf, and V. L. Jentsch, "Delayed effects of acute stress on cognitive emotion regulation," *Psychoneuroendocrinology*, vol. 125, p. 105101, 2021.

[67] Y. S. Can, B. Mahesh, and E. André, "Approaches, applications, and challenges in physiological emotion recognition - A tutorial overview," *Proc. IEEE*, vol. 111, no. 10, pp. 1287–1313, 2023. [Online]. Available: https://doi.org/10.1109/JPROC.2023.3286445

[68] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54 776–54 788, 2020.

[69] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, "Stress detection by machine learning and wearable sensors," in *26th International Conference on Intelligent User Interfaces-Companion*, 2021, pp. 43–45.

[70] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Transactions on affective computing*, vol. 9, no. 2, pp. 227–239, 2016.

[71] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020, pp. 51–57.

[72] D. Bajpai and L. He, "Evaluating knn performance on wesad dataset," in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2020, pp. 60–62.

[73] P. Zhang, F. Li, R. Zhao, R. Zhou, L. Du, Z. Zhao, X. Chen, and Z. Fang, "Real-time psychological stress detection according to ecg using deep learning," *Applied Sciences*, vol. 11, no. 9, p. 3838, 2021.

[74] M. Albaladejo-González, J. A. Ruipérez-Valiente, and F. Gómez Mármol, "Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 11 011–11 021, 2023.

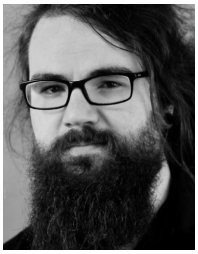
[75] P. Prajod, B. Mahesh, and E. André, "Stressor type matters!– exploring factors influencing cross-dataset generalizability of physiological stress detection," *arXiv preprint arXiv:2405.09563*, 2024.

[76] S. Gedam and S. Paul, "Automatic stress detection using wearable sensors and machine learning: A review," in *2020 11th International*

Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1–7.
[77] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: a meta-analysis and review of the literature," *Psychiatry investigation*, vol. 15, no. 3, p. 235, 2018.
[78] A. Greco, G. Valenza, J. Lázaro, J. M. G. Rey, J. Aguiló, C. de la Cámara, R. Bailón, and E. P. Scilingo, "Acute stress state classification based on electrodermal activity modeling," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 788–799, 2023. [Online]. Available: https://doi.org/10.1109/TAFFC.2021.3055294
[79] A. L. Callara, L. Sebastiani, N. Vanello, E. P. Scilingo, and A. Greco, "Parasympathetic-sympathetic causal interactions assessed by time-varying multivariate autoregressive modeling of electrodermal activity and heart-rate-variability," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 10, pp. 3019–3028, 2021.
[80] P. Gebhard, T. Schneeberger, T. Baur, and E. André, "Marssi: Model of appraisal, regulation, and social signal interpretation," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '18. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 497–506.
[81] P. Gebhard, T. Schneeberger, E. André, T. Baur, I. Damian, G. Mehlmann, C. König, and M. Langer, "Serious games for training social skills in job interviews," *IEEE Transactions on Games*, vol. 11, no. 4, pp. 340–351, 2019.
[82] A. Heimerl, S. Mertes, T. Schneeberger, T. Baur, A. Liu, L. Becker, N. Rohleder, P. Gebhard, and E. André, ""gan i hire you?" – a system for personalized virtual job interview training," 2022.

Alexander Heimerl obtained a master in Computer Science and Information Technology at the University of Augsburg, Germany. He is currently working as a research associate at the Lab for Human-Centered AI at the University of Augsburg. His research focuses on machine learning in the context of affective computing and explainable AI.

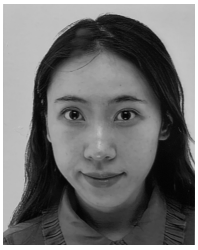
Pooja Prajod obtained a Masters degree in Computer Science from Delft University of Technology, Netherlands. She is currently a research associate at the Lab for Human-Centered AI at the University of Augsburg, Germany. Her research focuses on affective computing in human-robot collaboration.



Silvan Mertes obtained his master's degree in Computer Science at the University of Augsburg, Germany. He is currently working as a research associate at the Lab for Human-Centered AI at the University of Augsburg. His research focuses on generative models, especially in the context of explainable AI.

Tobias Baur is a post-doctoral researcher at the Human-Centered AI Lab, Augsburg University, Augsburg, Germany. He received his PhD (Doktor rer. nat.) in 2018. His main research focuses on human-centered tools that help (non-)experts to create and understand Machine Learning models. His research topics include Artificial Emotional Intelligence; Social Signal Processing and Ethical & Explainable AI.

Matthias Kraus is a post-doctoral researcher at the Human-Centered AI Lab, Augsburg University, Augsburg, Germany. He received his PhD (Dr. -Ing.) in 2022. His main research focuses on socially-aware conversational AI. His research topics include proactive dialogue systems, human-AI trust, and human-robot interaction.



Ailin Liu Ailin Liu received her bachelor's degree from Jacobs University Bremen, in 2021, and she is pursuing her master's degree in Data Science at RWTH Aachen University. She is currently working as a student research assistant at the Lab for Human-Centered AI at the University of Augsburg. Her research interests include human-computer interaction, affective computing, and applied machine learning.

Helen Risack Helen Risack obtained her bachelor's degree in psychology at the Döpfer University of Applied Sciences in Regensburg. She is currently in her psychology master's program with a focus on Clinical Psychology, Psychotherapy and Health at Friedrich-Schiller-University in Jena. Besides, she is working as a research assistant at the Chair of Human-Centered Artificial Intelligence at the University of Augsburg in Germany.

Nicolas Rohleder is a Professor at the University of Erlangen-Nürnberg and is leading the Chair of Health Psychology. His research interests include among other topics the impact of cumulative stress events on peripheral inflammatory processes across the lifespan, the health implications of trauma, depression, and chronic stress, and the determinants of psychological and biological stress responses. Throughout his career, he has received multiple awards, including the Herbert Weiner Early Career Award from the American Psychosomatic Society (APS) in 2013, the Curt P. Richter Award from the International Society of Psychoneuroendocrinology (PNEC) in 2011, and the Young Investigator Award from the American Federation for Aging Research (AFAR) in 2009. Further, he holds the position of President-elect of the International Society of Psychoneuroendocrinology (ISPNE) and serves on the Steering Committee of the German Endocrine Brain Immune Networks (GEBIN).

Elisabeth André is a full professor of Computer Science and Founding Chair of Human-Centered Artificial Intelligence at Augsburg University in Germany. She has a long track record in multimodal human-machine interaction, embodied conversational agents, social robotics, affective computing and social signal processing. Her work has won many awards including the Gottfried Wilhelm Leibniz Prize, the most important research funding award in Germany, and she is a member of the prestigious Academy of Europe, the German Academy of Sciences Leopoldina and the CHI Academy. In 2013, she was awarded a EurAI fellowship (European Association for Artificial Intelligence). In 2019, she was named one of the 10 most influential figures in the history of AI in Germany by National Society for Informatics (GI). From 2019 - 2022, she served as the Editor-in-Chief of IEEE Transactions on Affective Computing.

Linda Becker is a professor for General Psychology at Vinzenz Pallotti University (Vallendar, Germany). She received her master degrees in physics and psychology from the University of Bremen, Germany, in 2007 and 2012, respectively, and the Ph.D. degree in psychology from the Friedrich-Alexander University of Erlangen-Nürnberg (FAU), Germany, in 2016. In 2022, she finished her habilitation thesis entitled “Biopsychological investigation of the associations between stress and cognitive functioning and their relation with mental and physical health”, for which she was awarded from the FAU in 2023. Her research focuses on the associations between peripheral physiological and cognitive processes. Her main research interests are the associations between biological stress mechanisms and cognitive functioning and the health benefits of physical activity.