2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)

# Automatic Detection of Subjective, Annotated and Physiological Stress Responses from Video Data

Matthias Norden
*CITEC – Center for Cognitive Interaction Technology*
*Bielefeld University*
Bielefeld, Germany
mnorden@techfak.uni-bielefeld.de

Oliver T. Wolf
*Department of Cognitive Psychology*
*Institute of Cognitive Neuroscience*
*Ruhr University Bochum*
Bochum, Germany
oliver.t.wolf@ruhr-uni-bochum.de

Lennart Lehmann
*Digital Health Center*
*Hasso Plattner Institute*
*University of Potsdam*
Potsdam, Germany
lennart.lehmann@gmail.com

Katja Langer
*Department of Cognitive Psychology*
*Institute of Cognitive Neuroscience*
*Ruhr University Bochum*
Bochum, Germany
katja.langer@ruhr-uni-bochum.de

Christoph Lippert
*Digital Health Center*
*Hasso Plattner Institute*
*University of Potsdam*
Potsdam, Germany
christoph.lippert@hpi.de

Hanna Drimalla
*CITEC – Center for Cognitive Interaction Technology*
*Bielefeld University*
Bielefeld, Germany
drimalla@techfak.uni-bielefeld.de

***Abstract*—Machine-learning-based stress detection systems differ with respect to the *ground truth* used for training the algorithms. It is unclear how models trained on different facets of the stress reaction (e.g., biological, psychological, social) can be compared, interpreted and applied. In this study, we investigate the influence of the stress label on the performance of machine learning models trained on either vocal characteristics or facial expressions extracted from videos. We collected videos from 40 male participants while being exposed to the Trier Social Stress Test (TSST) and assessed self-reported, live observed, video-annotated and neuro-endocrinological stress levels. We train three standard machine learning models to separately predict different stress labels using either voice or facial cues. Analyzing the relationships of different stress facets we found that observers' annotations were significantly positively associated (live vs. video annotated, $\rho_s$ = .53). Similarly, the neuro-endocrinological stress indices correlated with each other (cortisol vs. sAA , $\rho_s$ = .39). Machine learning experiments resulted in predictions that were positively associated with panel-annotated stress levels showing significantly stronger correlations in voice-based models ($\rho_s$ = .54 vs. $\rho_s$ = .30). Predictions of self-reported stress were positively related to ground truth values for face-based ($\rho_s$ = .24) but not for voice-based models. There was no evidence for successful predictions of video-annotations or endocrinological stress levels in both settings. We provide evidence that machine learning models trained on different stress assessments perform differently and should be interpreted and applied accordingly. Implications and recommendations for future work on video-based stress detection are discussed.**

***Index Terms*—stress detection, facial expressions, voice, video, TSST, machine learning**

## I. Introduction

Advances in computational technologies and data availability have inspired approaches to automatically detect emotional states [1], [2], including psychological distress [3], [4]. Systems detecting stress aim to do so unobtrusively based on the analysis of verbal and non-verbal behavior in video and/or audio data, e.g., for monitoring a driver's stress level [5] or helping individuals to reduce their daily stress [6].

The application potential of such systems relies on the properties and quality of the underlying datasets, highlighting the relevance of a concrete stress definition, reliable stress induction and valid assessment of stress states. In particular, the "stress label" used as *ground truth* for training machine learning algorithms determines how the predictions of such a system can be interpreted. Importantly, the definition and respective assessment of this stress label differs based on one's perspective [7]–[9].

Most stakeholders from the health sector focus on the *biological* stress dimensions (i.e., effects on the cardiovascular, gastrointestinal and musculoskeletal systems). Many psychologists take into account the negative relationship between the *subjective feeling* of being stressed and psychological well-being [10], [11]. In interaction research, the detection expressed acute stress as a *social signal* to the interaction partner can be considered the most important aspect to focus on when investigating stress [12], [13].

Despite these conceptually important differences, most studies aiming for the development of automated stress detection systems, typically assess the stress label using either physiological measurements (e.g., saliva cortisol, heart rate, breathing rate) **or** subjective self-reports (i.e., stress questionnaires) **or** subjective ratings from external observers. Nevertheless, the final systems are often interpreted and applied from a general stress detection perspective rather than explicitly pointing to the stress dimension underlying the algorithm development. It is unclear if systems trained on different facets of stress can be applied equally. As a potential consequence, stress detection algorithms based on externally annotated videos may not be valid for predicting self-perceived stress in daily life nor for the biological consequences of acute stress.

In this study, we investigate the influence of the underlying *ground truth* for developing machine learning based stress detection frameworks using video data. This paper has three main contributions: First, we build an audiovisual dataset of the gold standard laboratory stress paradigm the Trier Social Stress Test (TSST) [14]–[16] by assessing conceptually different stress responses of the

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

participants. Second, we evaluate the performance of standard machine learning methods, depending on whether subjectively perceived, externally observed or the neuro-endocrinological stress responses are used as ground truth and compare all analyses for audio and visual modality. Third, we discuss implications for future studies on video-based stress detection.

## II. Related Work

### A. Ground Truth in Stress Videos

Most video-based stress detection frameworks use videos obtained from experimental stress induction settings to train machine learning models, e.g., deep neural networks [3], [17]–[19]. The majority of these studies incorporate classification designs in which the final ground truth is based on the assumption of successful stress induction in the experimental versus a control group rather than direct stress assessments (e.g., physiological measurements) [3]. Although this might be plausible for a general stress state detection, the underlying stress dimension (e.g., subjective, observed or biological) that the models learn remains unclear. Previous work has outlined the complexity of different systems, such as the sympathetic nervous system (SNS) and hypothalamus-pituitary-adrenal axis (HPA), and cognitive-emotional experiences, involved in the stress response and their interplay [20], [21]. A recent review of 36 studies [9] found a significant positive association between self-reported stress and cardiovascular measure in 28% of studies but also highlighted the heterogeneity in measures and methods as a limitation for comparability. Similarly, a prior review on the psychological and physiological responses to the Trier Social Stress Test reported a significant positive association between cortisol and perceived emotional stress variables in 25% of the studies [22]. Taken together, this implies that stress detection frameworks based on stress induction experiments cannot be interpreted in a general way but should be considered in dependence of its ground truth dimension of stress.

In many studies on machine-learning-based affect detection, external annotation methods are used as the underlying ground truth. A recent review investigating biases in 130 audiovisual datasets for emotion recognition showed that emotion labels were derived from external annotators in the majority of studies [23]. Similarly, stress levels of participants in videos have been externally annotated with different methods. For example, in [24] two annotators rated the participants' stress levels in videos of a stressful driving scenario whereas Aigrain and Spodenkowiecz used a crowd-sourcing platform for annotating the stress levels in their video dataset [7], [8]. Because of inconsistencies in the composition of annotators (e.g. gender, number, age), used questionnaires and rating aggregation across studies [23], the quality of the ratings and the interrater-agreement should be considered when developing, interpreting and applying stress detection models. Furthermore, there can be large differences between self- and observer-ratings, as shown for emotional experience [25], sleepiness [26] and reported stress levels [7], [8]. Taking the divergence between subjective, biological and observer-rated stress responses [27], it remains elusive which kind of stress is actually detected by machine learning models trained on videos of stressed people.

### B. Multidimensional Stress Detection in Videos

Several studies have applied classical machine learning methods (incl. deep neural networks) to detect stress in videos using combinations of facial [17], [18], [28]–[31], posture [8], [32], linguistic [33] and voice [19], [34] data. However, empirical work comparing assessments of conceptually different stress dimensions as the underlying ground truth for stress detection models is scarce. In two studies [7], [8], the authors acquired a video dataset where participants annotated their stress levels after watching their own videos, videos were externally annotated using the crowd-sourcing platform AMT and experts rated the physiological stress level based on heart rate variability (HRV). Aigrain *et al.* (2018) applied several SVMs for classification of different stress dimensions and compared behavioral with physiological features regarding their classification performance for different stress dimensions [8]. Spodenkowiecz *et al.* (2018) later used the same data and methods to compare self versus externally annotated stress [7]. While both studies analyzed the interplay of different stress dimensions and differences in predictive physiological and behavioral features, the influence of the underlying stress label on the model performances was not investigated. Moreover, the continuous stress labels were transformed to binary stress classes using self-defined thresholds which diminishes the informative value for the different stress dimensions. Heart rate variability used as the physiological stress label in this study can be an important indicator of the SNS involved in the stress response but is not covering the HPA system. Additionally, both studies focused on visual and sensor-derived physiological features, leaving the potentials of voice data for stress prediction open.

The recently published MuSe challenge dataset [35] of TSST videos annotated with externally annotated emotional labels (valence, arousal) and various physiological measurements (heart rate and respiration) has been used for the prediction of emotional states and an EDA-derived stress label. While in this dataset various physiological and externally annotated stress labels are available, information on the subjective and neuro-endocrine stress responses of the participants is missing. Baird *et al.* (2022) recently combined several datasets of the TSST (one being the MuSe dataset) and applied support vector regressors (SVR) as well as an LSTM architecture to predict physiological (e.g., cortisol levels, heart rate, respiration) and externally annotated emotional (e.g., valence and arousal) stress responses [19]. They modeled different dimensions of stress focusing on speech-derived and to some extent on facial features. However, analyses with respect to subjectively perceived stress as well as more pronounced facial feature analyses are not part of this work.

This overview shows the need for a holistic analysis combining exploration and comparisons of conceptually different stress dimensions and their influence on stress detection modeling across video modalities. Meta-analysis comparing highly differing datasets (e.g., different ground truth labels) and methods (e.g., video processing, feature extraction, modeling steps and evaluation reports) may lead to inconclusive interpretations across studies. In the following, we describe our methods to tackle the open question of how classical machine learning model performances and predictive features differ when trained on different stress labels on either facial or voice data extracted from the same underlying dataset.

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

## III. METHODS

To acquire a video dataset capturing stress-related non-verbal behavior, we recorded participants undertaking the gold standard stress induction paradigm Trier Social Stress Test (TSST [15], [16], [36]). For each participant, we assessed the subjective (i.e., self-reported stress levels), externally observed (i.e., live vs. post-hoc annotated stress levels) and two neuro-endocrine dimensions (i.e., saliva cortisol and alpha amylase) of the acute stress responses. Finally, we use this dataset to train separate standard machine learning models for the prediction of subjective, observed and neuro-endocrine stress. We compare the different model performances across facial and voice data extracted from the videos. In this concept paper, we do not aim to improve the state-of-the-art stress detection methods. In contrast, we use frequently used methods to analyze the role of different stress ground truth for detection of stress in video recordings of a standardized stress test. Code will be publicly available via GitHub.

### A. Data Collection

Forty healthy participants were recruited for the study. As commonly done in stress studies, we only recruited male participants to exclude confounding factors resulting from general sex differences and menstrual cycle effects on cortisol and emotional responsivity [37], [38]. Three participants could not complete all measurements, leaving 37 participants (M = 24.2 years, SD = 3.8 years) for the final evaluation. Each participant performed a slightly modified version of the Trier Social Stress Test (see [39] for general procedure). Participants were instructed to prepare a five minute speech to be presented in front of a mixed panel (one male, one female) dressed in white coats. Panel members were German speaking working students trained to not give or show any feedback during the whole procedure. After giving the interview, participants were asked to perform a mental arithmetic task for another three minutes and were prompted to start over when failing. During the whole procedure, participants were filmed via a camera placed behind the panel. Participants were able to see their own live recordings on a TV screen next to the panel. After finishing the stress induction procedure, the stress assessments (see next section) were conducted and participants were subsequently debriefed. The local ethics committee of the Faculty of Psychology at Ruhr University Bochum approved the experimental protocol and each volunteer signed an informed consent beforehand.

### B. Stress Assessments

To evaluate which dimensions of stress can be predicted using the stress test videos, we obtained several stress measurements (subjective, observed, neuro-endocrine).

**Subjective Stress:** Participants were asked to rate the amount of stress they perceived during the stress test on a standard visual analogue scale (VAS) [40] directly after finishing the stress test. Participants indicated their stress level ranging from 0 (“not at all stressful”) to 100 (“extremely stressful”).

**Externally Observed Stress**: Similarly, both panel members (male and female, between 20 and 30 years old) were asked to separately rate the amount of stress they believed the participant had perceived during the stress test. We used the mean value of both evaluations as the label for *live observed* stress. Additionally, four German speaking researchers (two female, aged 22-31 years) individually rated the participants’ stress levels after watching the video recordings in randomized order. To get used to the task, all researchers first rated four similar videos which are not part of the current dataset. This resembles a typical data annotation procedure for machine learning and enables the investigation of rater coherence as well as the exploration of the differences between live observed and video observed stress levels. We used the mean value of the four ratings as the label for the *externally annotated* stress label as commonly done in other studies using videos with externally annotated emotions [41].

**Neuro-endocrinological Stress:** For the evaluation of physiological stress, we obtained measurements of saliva cortisol concentrations and alpha amylase (sAA). Salivary cortisol is known to be a strong indicator for biological stress, reflecting the activation of the HPA system and is used for physiological stress evaluation [42]. Alpha amylase levels have been discussed to indicate sympathetic activity associated with psychosocial stress and has shown faster increases compared to the cortisol response [43], [44]. Four measurements were taken: 2 minutes before the TSST (baseline) as well as 2 minutes, 15 minutes and 55 minutes after completing the TSST. Following the findings from Miller [45], we used the peak reactivity (peak level minus baseline level) for measuring cortisol reactivity and applied the same index to sAA.

### C. Preprocessing, Feature Extraction and Transformation

We recorded a video of each participant (shoulder close-up) performing the TSST including the voices of participants and panel members. As the interview parts reflect more natural situations and less interruptions from the panel, we only used this part for our video analysis. Mean duration of the interview parts was 305s (SD = 20s), recorded with 25 frames per second. We used standard frame-based feature extraction methods on video and audio data separately, and transformed the extracted features to reduce the computational cost and capture the time-series property of the dataset.

**Audio Data:** We first extracted the audio data from the video files and then performed preprocessing: We converted the signal to a mono, 16 khz, 16 bit uncompressed signal and normalized it to the same volume level as done in [19] and [34]. In order to distinguish spoken parts of the participant from spoken parts of panel members and non-spoken parts, we used the python module inaSpeechSegmenter [46] and segmented the preprocessed audio signal into 1) participant speaking, 2) panel speaking, and 3) no speaking activity. As we were interested in the prediction of stress from non-verbal voice features of the participant, we excluded panel interventions and reconnected the remaining segments for low-level voice feature extraction.

We used the open source toolkit openSMILE to extract the eGeMAPS feature set [47] which has been used for similar tasks, including cortisol level and emotion prediction [19], [34], [35]. The eGeMAPS feature set contains 88 functional features (e.g., mean pitch, mean shimmer) and has been extracted for every participant’s processed.

**Facial Video Data:** To predict stress from facial features, we used the open source toolkit OpenFace 2.2.0 [48] to extract facial action unit features as previously done for similar tasks [35], [49]. Facial action units (AU) are obtained

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

by integrating abstract facial movements into 44 observable muscle movements (e.g. AU 10: upper lip raiser) using the Facial Action Coding System (FACS). These are widely used in behavioral science and automatic facial expression analysis. OpenFace extracts the presence (binary) as well as the intensity in the range [0,5] of 18 AU (except for AU28 for which only presence is extracted) from each video frame. For every AU, we computed several feature functionals that summarize the information of facial activity over the frame series data. We computed all features separately for parts of the video where the participant is speaking and where the participant is not speaking using automatic speech segmentation [46]. Specifically we used the *mean, standard deviation, kurtosis, skew* and *Shannon entropy* of the OpenFace intensity features and the *number of activations* of the binary features over the whole interview for silent and for speaking parts. Additionally, we computed the *standard deviation of time intervals between adjacent peaks of the* intensity features for the complete video. Overall, this resulted in 223 features per video.

### D. Machine Learning Models

We used classical machine learning models that have been applied previously for similar tasks [3]: For each modality (voice and facial expression), we apply regression tasks on the different stress dimension labels using separately trained support vector regressors (SVR), random forest regressors (RFR) and elastic nets (EN), as well as standard baseline models (dummy regressors using mean or random prediction strategies). SVRs were tuned using grid search with nested cross validation (see section below and [50]) for the choice of the kernel (linear, radial basis function, polynomial; used in [7], [8], [19]) with kernel coefficient gamma = $(n_{features} * var(X))^{-1}$ for not-linear kernels and degree = 3 for polynomial kernel, and regularization parameters (Ɛ = [0.001,0.01,0.1,0, 1, 10], C = [0.001, 0.01, 0.1, 0.5, 1]). For RFR, we tuned the maximal depth of trees (d = [2, 8, 16, 32]) and the minimum number of samples per leaf (n = [1, 2, 4, 8]) in a forest with 1000 trees accordingly. ElasticNets were tuned for the regularization penalties (l1/l2 ratio = [0, 0.5, 1]). All models were tuned separately for each modality and label using the same grids, to ensure fair comparisons. We z-standardized the features within the model pipeline. In order to be able to compare the models across stress dimensions, we used scaling-invariant evaluation metrics (see section below).

**Evaluation Process:** To obtain an overall robust performance evaluation of the different models across modalities and labels, we applied nested cross validation with a 10-fold outer and a 5-fold inner loop as suggested in [50]. The differentiation between higher and lower stress with respect to a certain stress dimension can be seen as the primary goal of this study and allows for better comparison across potentially differently distributed and scaled stress labels. Therefore, we calculate Spearman's rank correlation coefficient ($\rho_s$) between the predicted values and the true values and report the mean value after 100 iterations as the overall performance metric for each model and stress label. We report only on average positive associations, as we consider negative correlation coefficients as artifacts of the cross validation scheme resulting from models not learning predictive information as described in [51]. To further compare the performance of models yielding positively correlated predictions across dimensions, we computed the percentage decrease from mean absolute error of the two baseline (random and mean) regressors' MAE separately and tested for significance using one-sample t-tests. In order to compare differences in performances between modalities, we analyzed the mean correlation coefficients over all iterations using the non-parametric Mann-Whitney-U-test.

## IV. Results

In this section, we describe the acquired dataset, including inter-rater agreements, distribution and correlation of the different stress labels over all participants, and report the voice feature-based and facial feature-based model performances.

### A. Stress Assessments (Ground Truth)

To assess the inter-rater agreement, Spearman correlation between the separate ratings was calculated showing strong agreement in the panel members' ranking of the participants' stress levels ($\rho_s = .75$ , $p < .001$) and low agreement between external raters (mean pairwise $\rho_s = .23$). Descriptive statistics of the participants' stress levels according to the five different stress assessments are given in Table I. Stress assessments highly differed both, in scale and distributions with subjective stress ratings (self-reported, panel and video-annotated) being mainly left-skewed whereas neuro-endocrinological stress indices are right-skewed. The observed stress labels (live panel annotated and video annotated) were associated positively ($\rho_s = .53$, $p < .001$), as well as the neuro-endocrinological stress labels (cortisol and sAA , $\rho_s = .39$, $p = .018$). Self- and panel ratings were marginally significantly correlated ($\rho_s = .31$, $p = .061$). There was no evidence for other significant associations among the stress labels.

TABLE I. Descriptive statistics of the stress assessments

| | M | SD | Median | Min | Max | Skew |
|---|---|---|---|---|---|---|
| **SELF** | 67.0 | 22.9 | 80.0 | 14.0 | 100.0 | -0.8 |
| **PANEL** | 60.4 | 17.9 | 60.0 | 15.0 | 100.0 | -0.3 |
| **ANNOT** | 47.5 | 11.8 | 46.8 | 19.3 | 70.5 | -0.3 |
| **CORT** | 2.9 | 3.4 | 2.5 | 0.0 | 14.4 | 1.7 |
| **sAA** | 94.6 | 85.3 | 65.9 | 0.0 | 380.1 | 1.7 |

SELF, PANEL, ANNOT refer to self-reported, panel-annotated and video-annotated stress levels. CORT and sAA refer to cortisol and sAA peak reactivity stress indices.

### B. Model Performances

**Voice-based Prediction:** The average positive mean Spearman correlation coefficients over all iterations are shown in Fig. 1A. Voice-based models achieved best correlated predictions for the annotations of the live panel ($\rho_s = .54 \pm .04$ for SVR). The SVR model increased the random regressor baseline performance significantly by 73% (SD = 22%, $p < .001$) and the mean predicting baseline performance by 20% (SD = 6%, $p < .001$) with respect to MAE. Model predictions for self- reported, video annotated, sAA and cortisol stress levels were not strongly positively correlated (all $\rho_s < .01$) with respective ground truth values.

**Face-based Prediction:** The highest positive correlations between facial feature-based model predictions and ground truth labels were achieved for panel-rated ($\rho_s = .30 \pm .08$ for EN) and self-rated ($\rho_s = .27 \pm 0.1$ for EN) stress levels (see Fig. 1B). On average, optimized elastic nets increased the random regressor baseline model performance by 41% (SD = 17%, $p < .001$) w.r.t. MAE when predicting panel-rated and 28% (SD = 17%, $p < .001$) when predicting self-rated stress levels.

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

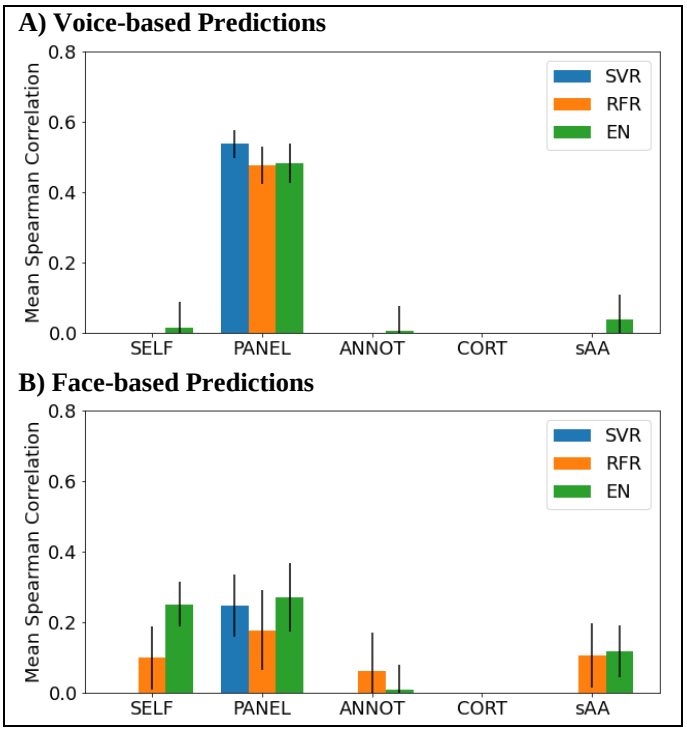


Fig. 1. Mean Spearman correlation coefficients between model predictions and ground truth values for the stress labels using voice (A) or facial (B) features extracted from the videos. For each iteration (n = 100), models were evaluated using 10-fold shuffled cross validation applying nested grid search. Error bars indicate standard deviation of the iteration results. Negative averaged correlations are set to zero, as they are representing predictions of models not learning any predictive information.

Both models did not substantially increase the mean predicting baseline performance regarding MAE (< 5%).

Similar to voice feature based models, predictions for self-reported, video annotated, sAA and cortisol stress levels were not strongly positively correlated (all $\rho_s < .02$) with respective ground truth values. The panel ratings could be predicted better with voice-based than with face-based models, evident in a Mann-Whitney-U test of the respective predictions' correlation with the ground truth ($U = 9985$, $p < .001$). In contrast, only face-based models were able to predict self-reported stress.

## V. Discussion

In this study, we explored conceptually different stress dimensions in an audiovisual dataset of TSST participants and investigated how the choice of the stress label might influence the predictive performance in machine learning pipelines. We found that video-based stress ratings were annotator-specific and showed that self-reported, observable and neuro-endocrinological stress assessments differed within participants. Voice and facial-feature-based models allowed a meaningful prediction of externally observable stress while self-reported stress levels were only partially predictable. We were not able to predict neuro-endocrinological stress levels. In the following, we discuss these results and tackle practical implications for future work.

### A. Stress Ratings

First, stress ratings showed heterogeneity with respect to inter-rater agreement and measurement type. Whereas there was high agreement between live panel raters, inter-rater agreement between video annotators was poor. Nevertheless, aggregated live panel and video annotator ratings were significantly correlated indicating an overall agreement within stress level observations based on observation. The participants' self-reported stress levels were marginally significantly associated with live panel ratings whereas there was no evidence for associations with video annotations. Differences between ratings within each annotation procedure might be caused by personal, cultural and social backgrounds of the annotators as well as situational aspects [52], [53]. Annotation procedures highly differ regarding "involvement" (i.e., live vs. video), composition (e.g., gender), data aggregation and reported statistical measures of inter-rater agreements [23]. The higher agreement between live-annotators might be caused by the possibility to align by exchanging information on participants' presentations. On the other hand, live panel members might perceive additional signals that are not observable in videos and also allow for capturing the participants' actual stress feeling. This is supported by previous work showing differences in emotion processing for personal and non-personal contact [54]. In our study, four behavioral researchers rated the videos in randomized orders after watching the same example videos. Their high disagreement indicates the heterogeneity in one's perception of affects, especially when transmitted via screens only.

Importantly, variances in annotations might lead to inconsistent and none reproducible results across studies. Using aggregated ratings from several raters with a high level of disagreement as the ground truth puts the validity for generalized application of such algorithms in question and points to the necessity of personalized approaches. With external video annotations still forming the lion's share in dataset labeling for affect recognition [23], we argue for using robust annotation procedures including a higher number of annotators and additional briefings. Annotators might be chosen according to a systems' future applications' target group (e.g., age-group, cultural group). Additionally, annotation results should be reported and discussed explicitly with respect to potential applications when developing stress detection systems.

In our experiments, model predictions of panel-rated stress levels were positively correlated with ground truth values both, for voice and facial-feature-based approaches but no meaningful associations were found for predictions of video-annotated stress levels. This supports our assumption that machine learning model performances differ with respect to the external annotations used as ground truth for observable stress. Despite predictive information for evaluating the external stress appearance captured in and detected from participants' non-verbal behavior, the high inter-rater disagreement between video annotators might lead to models that are not useful for different users. Comparing voice and facial feature based models, the former achieved significantly higher correlation coefficients in predicting the panel-rated stress levels. Similarly, [19] and [35] reported higher performances when using voice features for the prediction of externally annotated emotional labels, suggesting that the social signal of stress is mainly perceived via vocal indicators.

The predictions of the live panel-rated stress levels were significantly better correlated with ground truth values than all other predictions, including the prediction of self-reported stress levels, which were positively (but not significantly)

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

associated with ground truth values only for models using facial features. This is in line with [7] and [8] who found higher classification performances for predicting video annotated stress levels than for predicting self-reported stress from non-verbal behavioral features. We find that models seem to perform better for predicting annotations based on observable behavior than for self-reported ratings. This might be attributed to the fact that the models actually "see" the same non-verbal behavioral clues as human observers and struggle the same way when predicting actual perceived stress. This has some important implications for applying such algorithms. For example, a driver-monitoring system developed based on external annotations as in [24] might be able to detect the driver *appearing* stressed (similarly as a co-driver might evaluate) but the actual underlying (subjective and physiological) stress state might still be overlooked and potential interventions (e.g., driver alarms) mistargeted. In the same way, stress prevention or treatments based on automatic stress detection might just detect the outer appearance of being stressed and therefore not be beneficial for its user. On the other hand, emotional suppression [55] might lead to systems overlooking users' stress and misplacing possible interventions.

### B. Neuro-endocrinological Stress

We also assessed the predictability of two important biomarkers of stress using facial or voice features extracted from stress test videos. In our experimental dataset cortisol and sAA indices were significantly correlated but did not show significant associations to other stress assessments (self-reported, observed stress levels). This is in line with some previous works on the interplay of physiological and subjective stress [21], [27], [56] and on studies of multidimensional stress detection [8], whereas other studies did find significant associations [57], [58].

While in our study model predictions for live panel-rated stress levels were positively correlated with ground truth values and self-ratings could be partially predicted, predictive performance for the neuro-endocrinological stress labels was poor across both modalities. This implies that in the same way human stress evaluations differ from actual neuro-endocrinological processes, models did not learn meaningful information from the voice and facial features. Systems promising to detect stress aspects relevant for (mental) health (e.g., release of stress hormones, blood pressure changes or stress feelings) in videos but developed using external stress annotation should be seen critically.

Although associations between facial [32] and voice [59] characteristics with respect to biological stress markers, including cortisol, were found in previous work and additionally some studies used machine learning models for predicting such markers [19], [34], only using non-verbal behavioral features from video recordings seems to be difficult. Baird *et al*. showed that correlations differ with respect to the time point of cortisol measurements with highest correlations between model predictions based on TSST voice samples and cortisol values ($\rho_s$ = .421) 20 minutes after the stress test [34]. In [19], they achieved correlations of $\rho_s$ = .770 based on voice samples using LSTMs, suggesting that sampling methods and more complex models might be able to capture physiological stress relevant information. On the other hand, their results might also be influenced by gender-dependent stress and voice characteristics. Additionally, we transformed the raw cortisol values to a stress index as recommended by [45] and did not predict cortisol levels at different time points which impedes a direct comparison of the study results.

Aigrain *et al*. showed that including physiological features improved model performances for classifying HRV-derived biological stress [8]. Indeed, many studies incorporate sensor-derived physiological signals to predict stress and achieve better results [4]. Based on the findings of these previous studies and the results of our work obtained with standard methods and robust cross validation evaluation schemes, we conclude that detecting biological consequences of acute stress using only facial and voice features from videos needs to be questioned. Including other potentially stress-relevant cues (e.g., head-pose, posture, eye-gaze, linguistic features) from the videos, applying feature selection methods, and using more complex models might improve the results. Still, the health promises of current systems should be viewed with caution.

### C. Limitations and Future Work

Many stress detection studies are limited with respect to the quality and size of datasets [3], [19]. We exceeded several previous attempts [7], [8], [60] and collected videos capturing each five minutes of 37 participants undergoing the gold standard stress induction paradigm TSST including continuous and directly assessed stress markers instead of mere "stress vs. non-stress" classification labels [7], [18], [28]. Still, bigger sample sizes or data augmentation methods (e.g., sampling) and an additional control group may improve the stress detection. Nevertheless, we applied best practices for working with small datasets and reported conservative and robust evaluation metrics across the stress dimensions [50].

The stress assessments we obtained in this study have been used previously as ground truth for building stress detection models [3], [8], [19], [27], [34]. While these assessments reflect important aspects of the complex processes underlying the stress responses, some fine-grained information is not focused on in this study. For example, the SNS assessment through sAA might be enhanced by measurements of the heart rate and heart rate variability. Subjective evaluations (self-reports and annotations) as well as biological stress consequences differ from moment to moment. Thus, in future work, we plan to include additional continuous assessments and time-dependent analyses. We also pointed to the difficulties when using annotators' aggregated evaluations as the ground truth of observable stress. Insights from studies that investigate the relationships between annotator composition, aggregation scheme, inter-rater agreement and model performances could help to design more robust and real-world applicable models in the future.

## VI. Conclusion

Video-based stress detection is one key focus in the affective computing domain and carries potentials to unobtrusively monitor drivers' stress or improve daily stress prevention programs. Most studies differ with respect to the methods and the underlying data that are used for developing such algorithms. In this paper, we explored different stress dimensions and investigated how the choice of the stress label might influence the predictive performance in machine learning pipelines. We showed that among subjective, externally annotated and neuro-endocrinological stress labels,

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

model predictions were only positively associated with ground truth for annotated stress labels. Few or no predictive information were learned by models trained on neuro-endocrinological stress indices. Based on the findings, we conclude three main recommendations for future studies on video-based stress detection using machine learning methods:

First, researchers should report and interpret results of video-based stress detection models keeping in mind the underlying ground truth stress dimension used for algorithm development. This includes explicit discussions on the application potential of such algorithms, in particular when it's not clear which stress dimension is actually detected (i.e., no direct stress assessment). In this case, we suggest to (re-)evaluate the algorithm regarding the stress dimension of interest. Secondly, we advise future studies to separately investigate, report and interpret model performances with respect to *more than one* stress dimension. This comprises conducting experiments with direct assessments of *multiple* stress dimensions (subjective reports, observer reports, physiological markers) repeatedly over the experiments. Lastly, we opt for keeping the quality and reliability of the underlying dataset as a key aspect when developing algorithms. Optimally, experiments should follow literature-proven protocols and measurements. Additionally, annotators should be chosen with the specific use-case in mind, and specifics of the annotation process and inter-rater disagreements evaluated, reported and discussed.

## Ethical Impact Statement

Despite its beneficial application potentials, automated and unobtrusive stress detection systems should be discussed critically. Video recordings might be analyzed without prior consent or even without individuals being aware of it, especially in the context of surveillance. Sensitive information on individuals' mental health might be exploited and used for deceptive purposes as for hiring decision or insurance policy making. Systems applied for beneficial purposes such as driver monitoring, must work reliably, deliver valid assessments and users must be aware of a systems' limitations.

In our work, we did not focus on improving stress detection algorithms or specific applications. In contrast, we highlighted the importance and implications of the underlying stress dimension and argued for a re-evaluation of algorithms before applying systems to specific use cases. Promises of existing applications should be taken with caution. Similar to many other stress studies, our results need to be interpreted with the limited and biased sample (i.e., consisting of young male and single-cultured participants) in mind. We favored conservative evaluations but still advise that stress detection algorithms' performances might differ even more when applied with other genders and/or different cultures.

We hope that our study inspires future research groups to critically review stress datasets and carefully plan experiments before inducing stress in participants. Secondly, we advise potential users to take promises of (existing) applications with caution and be encouraged to inform themselves on a system's underlying ground truth. Lastly, we hope that our study motivates developers to grasp basic research before designing actual real world applications.

## Acknowledgment

This study was funded by the Federal Ministry of Education and Research Germany (BMBF; 01IS20046). The BMBF had no involvement in the study design, the collection, analysis and interpretation of data, the writing of the report and the decision to submit the article for publication.

## References

[1] J. Garcia-Garcia, V. Penichet, and M. Lozano, *Emotion detection: a technology review*. 2017, p. 8.

[2] D. Y. Liliana and T. Basaruddin, "Review of Automatic Emotion Recognition Through Facial Expression Analysis," Oct. 2018, pp. 231–236.

[3] T. A. Roldán-Rojo, E. Rendón-Veléz, and S. Carrizosa, "Stressors and Algorithms Used for Stress Detection: a Review," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2021, pp. 1–8.

[4] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, Jul. 2019.

[5] H. Gao, A. Yüce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 5961–5965.

[6] C. M. A. Ilyas, S. Song, and H. Gunes, "Inferring User Facial Affect in Work-like Settings," *arXiv:2111.11862 [cs]*, Nov. 2021. Accessed: Mar. 12, 2022.

[7] M. Spodenkiewicz, J. Aigrain, N. Bourvis, S. Dubuisson, M. Chetouani, and D. Cohen, "Distinguish self- and hetero-perceived stress through behavioral imaging and physiological features," *Prog Neuropsychopharmacol Biol Psychiatry*, vol. 82, pp. 107–114, Mar. 2018.

[8] J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani, "Multimodal Stress Detection from Multiple Assessments," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, vol. 9, no. 4, pp. 491–506, Oct. 2018.

[9] T. Vaessen *et al.*, "The association between self-reported stress and cardiovascular measures in daily life: A systematic review," *PLoS One*, vol. 16, no. 11, p. e0259557, 2021.

[10] O. Strizhitskaya*, M. Petrash, S. Savenysheva, I. Murtazina, and L. Golovey, "Perceived Stress And Psychological Well-Being: The Role Of The Emotional Stability," *European Proceedings of Social and Behavioural Sciences*, vol. Cognitive-Social, and Behavioural Sciences-icCSBs 2018, Feb. 2019. Accessed: Aug. 09, 2022.

[11] M. A. Griffin and S. Clarke, "Stress and well-being at work," in *APA handbook of industrial and organizational psychology, Vol 3: Maintaining, expanding, and contracting the organization*, Washington, DC, US: American Psychological Association, 2011, pp. 359–397.

[12] S. Vinanzi, A. Cangelosi, and C. Goerick, "The collaborative mind: intention reading and trust in human-robot interaction," *iScience*, vol. 24, no. 2, p. 102130, Feb. 2021.

[13] A. Vinciarelli and H. Salamin, "Social Signal Processing: Understanding Social Interactions through Nonverbal Behavior Analysis."

[14] A. P. Allen, P. J. Kennedy, S. Dockray, J. F. Cryan, T. G. Dinan, and G. Clarke, "The Trier Social Stress Test: Principles and practice," *Neurobiology of Stress*, vol. 6, pp. 113–126, Feb. 2017.

[15] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test'--a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1–2, pp. 76–81, 1993.

[16] B. M. Kudielka, D. H. Hellhammer, and C. Kirschbaum, *Ten years of research with the Trier Social Stress Test (TSST) – revisited*.

[17] H. Zhang, L. Feng, N. Li, Z. Jin, and L. Cao, "Video-Based Stress Detection through Deep Learning," *Sensors*, vol. 20, no. 19, p. 5552, Sep. 2020.

[18] G. Giannakakis *et al.*, "Stress and anxiety detection using facial cues from videos," *BIOMEDICAL SIGNAL PROCESSING AND CONTROL*, vol. 31, pp. 89–101, Jan. 2017.

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

[19] A. Baird *et al.*, "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress," *Frontiers in Computer Science*, vol. 3, 2021. Accessed: Feb. 16, 2022.

[20] M. Joëls and T. Z. Baram, "The neuro-symphony of stress," *Nat Rev Neurosci*, vol. 10, no. 6, pp. 459–466, Jun. 2009.

[21] J. Campbell and U. Ehlert, "Acute psychosocial stress: Does the emotional stress response correspond with physiological responses?," *Psychoneuroendocrinology*, vol. 37, no. 8, pp. 1111–1134, Aug. 2012.

[22] J. Andrews, N. Ali, and J. C. Pruessner, "Reflections on the interaction of psychogenic stress systems in humans: The stress coherence/compensation model," *Psychoneuroendocrinology*, vol. 38, no. 7, pp. 947–961, Jul. 2013.

[23] W. Saakyan, O. Hakobyan, and H. Drimalla, "Representational bias in expression and annotation of emotions in audiovisual databases," Bologna, Italy, 2021.

[24] J. Healey and R. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, pp. 156–166, Jul. 2005.

[25] K. P. Truong, M. A. Neerincx, and D. A. van Leeuwen, "Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 318–321.

[26] C. Ahlstrom, C. Fors, A. Anund, and D. Hallvig, "Video-based observer rated sleepiness versus self-reported subjective sleepiness in real road driving," *Eur. Transp. Res. Rev.*, vol. 7, no. 4, Art. no. 4, Dec. 2015.

[27] E. S. Epel *et al.*, "More than a feeling: A unified view of stress measurement for population science," *Front Neuroendocrinol*, vol. 49, pp. 146–169, Apr. 2018.

[28] 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)*, Nov. 2020, pp. 728–733.

[29] M. Gavrilescu and N. Vizireanu, "Predicting Depression, Anxiety, and Stress Levels from Videos Using the Facial Action Coding System.," *Sensors (Basel)*, vol. 19, no. 17, Aug. 2019.

[30] C. Viegas, S.-H. Lau, R. Maxion, and A. Hauptmann, "Towards independent stress detection: A dependent model using facial action units," in *2018 international conference on content-based multimedia indexing (CBMI)*, Sep. 2018, pp. 1–6.

[31] C. Viegas, S.-H. Lau, R. Maxion, and A. Hauptmann, "Distinction of stress and non-stress tasks using facial action units," in *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*, New York, NY, USA, Oct. 2018, pp. 1–6.

[32] 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)*, May 2015, vol. 03, pp. 1–6.

[33] I. Lefter, G. Burghouts, and L. Rothkrantz, "Recognizing Stress Using Semantics and Modulation of Speech and Gestures," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, vol. 7, no. 2, pp. 162–175, Apr. 2016.

[34] A. Baird *et al.*, "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," in *Interspeech 2019*, Sep. 2019, pp. 534–538.

[35] L. Stappen *et al.*, "MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox," *arXiv:2107.11757 [cs]*, Jul. 2021. Accessed: Aug. 04, 2021.

[36] M. A. Birkett, "The Trier Social Stress Test Protocol for Inducing Psychological Stress," *JoVE*, no. 56, p. 3238, Oct. 2011.

[37] J. J. W. Liu, N. Ein, K. Peck, V. Huang, J. C. Pruessner, and K. Vickers, "Sex differences in salivary cortisol reactivity to the Trier Social Stress Test (TSST): A meta-analysis," *Psychoneuroendocrinology*, vol. 82, pp. 26–37, Aug. 2017.

[38] P. M. Maki *et al.*, "Menstrual cycle effects on cortisol responsivity and emotional retrieval following a psychosocial stressor," *Horm Behav*, vol. 74, pp. 201–208, Aug. 2015.

[39] I. Labuschagne, C. Grace, P. Rendell, G. Terrett, and M. Heinrichs, "An introductory guide to conducting the Trier Social Stress Test," *Neurosci Biobehav Rev*, vol. 107, pp. 686–695, Dec. 2019.

[40] F.-X. Lesage, S. Berjot, and F. Deschamps, "Clinical stress assessment using a visual analogue scale," *Occup Med (Lond)*, vol. 62, no. 8, pp. 600–605, Dec. 2012.

[41] M. Valstar *et al.*, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*, Orlando, Florida, USA, 2014, pp. 3–10.

[42] B. Kudielka, D. Hellhammer, and C. Kirschbaum, "Ten years of research with the Trier Social Stress Test—revisited," in *Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior*, 2007, pp. 56–83.

[43] N. Ali and U. M. Nater, "Salivary Alpha-Amylase as a Biomarker of Stress in Behavioral Medicine," *Int.J. Behav. Med.*, vol. 27, no. 3, pp. 337–342, Jun. 2020.

[44] U. M. Nater and N. Rohleder, "Salivary alpha-amylase as a non-invasive biomarker for the sympathetic nervous system: current state of research," *Psychoneuroendocrinology*, vol. 34, no. 4, pp. 486–496, May 2009.

[45] R. Miller, F. Plessow, C. Kirschbaum, and T. Stalder, "Classification criteria for distinguishing cortisol responders from nonresponders to psychosocial stress: evaluation of salivary cortisol pulse detection in panel designs," *Psychosom Med*, vol. 75, no. 9, pp. 832–840, Dec. 2013.

[46] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "AN OPEN-SOURCE SPEAKER GENDER DETECTION FRAMEWORK FOR MONITORING GENDER EQUALITY," Calgary, Canada, Apr. 2018.

[47] F. Eyben *et al.*, "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, Apr. 2016.

[48] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–10.

[49] H. Drimalla, I. Baskow, B. Behnia, S. Roepke, and I. Dziobek, "Imitation and recognition of facial emotions in autism: a computer vision approach," *Molecular Autism*, vol. 12, no. 1, p. 27, Dec. 2021.

[50] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, p. e0224365, Nov. 2019.

[51] R. A. Poldrack, G. Huckins, and G. Varoquaux, "Establishment of Best Practices for Evidence for Prediction A Review," *JAMA Psychiatry*, vol. 77, no. 5, pp. 534–540, May 2020.

[52] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," *Psychol Sci Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019.

[53] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in Emotion Perception," *Curr Dir Psychol Sci*, vol. 20, no. 5, pp. 286–290, Oct. 2011.

[54] C.-T. Hsu, W. Sato, and S. Yoshikawa, "Enhanced emotional and motor responses to live versus videotaped dynamic facial expressions," *Sci Rep*, vol. 10, no. 1, Art. no. 1, Oct. 2020.

[55] J. J. Gross and R. W. Levenson, "Emotional suppression: physiology, self-report, and expressive behavior," *J Pers Soc Psychol*, vol. 64, no. 6, pp. 970–986, Jun. 1993.

[56] I. B. Mauss, R. W. Levenson, L. McCarter, F. H. Wilhelm, and J. J. Gross, "The tie that binds? Coherence among emotion experience, behavior, and physiology," *Emotion*, vol. 5, no. 2, pp. 175–190, Jun. 2005.

[57] U. Rimmele *et al.*, "Trained men show lower cortisol, heart rate and psychological responses to psychosocial stress compared with untrained men," *Psychoneuroendocrinology*, vol. 32, no. 6, pp. 627–635, Jul. 2007.

[58] W. Schlotz, R. Kumsta, I. Layes, S. Entringer, A. Jones, and S. Wüst, "Covariance between psychological and endocrine responses to pharmacological challenge and psychosocial stress: a question of timing," *Psychosom Med*, vol. 70, no. 7, pp. 787–796, Sep. 2008.

[59] K. Pisanski, J. Nowak, and P. Sorokowski, "Individual differences in cortisol stress response predict increases in voice pitch during exam stress.," *Physiol Behav*, vol. 163, pp. 234–238, Sep. 2016.

[60] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke, "Thermal spatio-temporal data for stress recognition," *EURASIP JOURNAL ON IMAGE AND VIDEO PROCESSING*, Jun. 2014.


978-1-6654-5908-2/22/$31.00 ©2022 IEEE