Machine Learning-Based Detection of Acute Psychosocial Stress from Digital Biomarkers

Master's Thesis in Medical Engineering

submitted
by

Victoria Müller

born 15.09.1996 in Freising

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

in Cooperation with

Department of Psychiatry, New York University Langone Health[1]
Chair of Health Psychology, FAU[2]

Advisors: Robert Richer M.Sc, Luca Abel M.Sc., Prof. Dr. Bjoern Eskofier,
Prof. Dr. Katharina Schultebraucks[1],
Prof. Dr. Nicolas Rohleder[2]

Started: 01.12.2023 Finished: 28.05.2024

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.
Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 28. Mai 2024

Übersicht

Die traditionelle Stressmessung, oft kostspielig und zeitaufwendig, kann das natürliche Verhalten in akuten Stresssituationen störend beeinflussen. Videoaufnahmen, als nicht-invasive und skalierbare Alternative, nutzen“digitale Biomarker” wie Gesichtsausdrücke, Kopfbewegungen und Blickverhalten zur Stresserkennung. Trotz vielversprechender Ergebnisse schränken kleine Stichprobengrößen und das Fehlen einer Kontrollbedingung die Generalisierbarkeit ein. Diese Thesis erforscht die Erkennung von akutem Stress durch digitale Biomarker und erweitert die Analyse um die videobasierte Herzratenmessung mittels Remote Photoplethysmography (rPPG) zur Modellierung physiologischer Veränderungen. Dabei wird die rPPG-basierte Herzrate unter dynamischen Bedingungen wie variabler Herzrate und Sprechaktivitäten untersucht und in einem multimodalen Ansatz mit digitalen Biomarkern für eine umfassende Stresserkennung kombiniert.

Vierundvierzig gesunde Individuen (57% Frauen) unterzogen sich dem Trier Social Stress Test (TSST) und dem friendly TSST (f-TSST), beide videoaufgezeichnet und bestehend aus einem Interview und einem Mathetest, in randomisierter Reihenfolge an zwei aufeinanderfolgenden Tagen. Die Analyse der digitalen Biomarker zeigte, dass während des TSST signifikant weniger positive Gesichtsausdrücke, reduzierte Kopfbewegungen und ein statischeres Blickverhalten gezeigt wurden als beim f-TSST. Für die Unterscheidung zwischen TSST und f-TSST basierend auf digitalen Biomarkern erzielten Modelle für maschinelles Lernen (ML) eine Accuracy von 73.3$\pm$5.5%. Explainable AI identifizierte eine Kombination aus Gesichtsausdrücken, Körperbewegungen und Blickverhalten als die zehn wichtigsten Marker. Im Matheteil waren ähnliche Marker relevant, hingegen wurden im Interview vor allem Gesichtsausdrücke für die Klassifikation genutzt.

Während die rPPG-Modelle auf den kontrollierten Benchmark-Datensätzen UBFC-rPPG und PURE vielversprechend waren, sank die Genauigkeit auf dem dynamischeren (f-)TSST, insbesondere bei vermehrter Körperbewegung, variabler HR und Sprachsequenzen. Die Integration der rPPG-basierten HR in die ML-Modelle verbesserte deren Robustheit, insbesondere während dem Matheteil, in der sich die Accuracy auf 77.3$\pm$6.5% verbesserte. Explainable AI identifizierte die rPPG-basierte HR als eine der drei einlfussreichsten Marker auf die Modelle.

Zusammenfassend betont diese Arbeit das Potenzial videobasierter digitaler Biomarker zur Erkennung von akutem psychosozialem Stress und zum besseren Verständnis von Stressreaktionen. Der Einfluss verschiedener Stressoren auf das Verhalten bedarf weiterer Forschung. Obwohl die rPPG-Modelle vielversprechend sind, müssen sie für eine zuverlässige Anwendung in der Praxis noch verfeinert werden. Zukünftige Forschungen sollten videobasierte Methoden weiterentwickeln und Stresserkennungstechnologien robuster und anwendbarer machen, um das Stressmanagement in klinischen, beruflichen und alltäglichen Umgebungen durch digitale Biomarker zu vereinfachen.

Abstract

Despite being well established, traditional stress measurement methods are typically costly and time consuming. Video recordings offer a scalable, contactless way to observe human behavior by extracting facial expressions, head movements, or gaze behavior – known as "digital biomarkers". Although initial findings for stress detection are promising, issues like small sample sizes and lack of control conditions highlight the need for more research. This thesis investigates the contactless detection of acute stress using video-based digital biomarkers. To enhance the understanding of the interaction between behavioral and physiological stress responses, this thesis evaluates remote photoplethysmography (rPPG) for estimating heart rate (HR) from facial videos under more realistic conditions such as variable HR and speech. Finally, by integrating rPPG-derived HR with digital biomarkers a multimodal approach is explored for comprehensive stress detection.

Forty-four healthy individuals (57% women) underwent the Trier Social Stress Test (TSST) and the friendly TSST (f-TSST), which consist of an interview (Talk) and a mental arithmetic test (Math), in randomized order on two consecutive days while being video recorded. Analysis of the extracted video-based digital biomarkers revealed that individuals displayed significantly less positive facial expressions, reduced head movements, and more static gaze behavior during the TSST compared to the f-TSST. Training machine learning models on these digital biomarkers yielded an accuracy of 73.3±5.5% in detecting exposure to the TSST or f-TSST. Explainable AI algorithms identified a mix of facial expressions, body movements, and pupil dynamics as top ten features. Notably, while the same set of features were influential during the Math phase, only facial expression features were the most important ones for decision-making in the Talk phase.

While rPPG models showed promise in controlled settings namley on the benchmark datasets UBFC-rPPG and PURE, when tested on the f-TSST, their accuracy decreased due to the more dynamic environment, particularly during tasks with high movement, variable HR, and speech. However, integrating rPPG-derived HR into the stress prediction models enhanced their robustness, particularly during the Math phase, where accuracy improved to 77.3±6.5%. Explainable AI identified rPPG-derived HR as one of the top three most influential features for the models.

In conclusion, this thesis underscores the potential of video-based digital biomarkers for detecting acute psychosocial stress and advancing the understanding of behavioral and physiological stress responses. It shows that stressors uniquely influence human behavior and emphasizes the need for further research. Although promising, rPPG models need refinement for reliable real-world use. Future work should enhance video-based methods and expand data collection to improve the robustness and applicability of stress detection technologies, leveraging digital biomarkers to simplify stress management across clinical, workplace, and everyday settings.

Contents

Chapter 1

Introduction

Stress, an ubiquitous presence in our daily lives, plays a dual role. While it is a natural bodily reaction to cope with challenges, excessive stress can heavily impact both physical health and mental well-being, leading to long-term sickness [OCo21], [APA19], [Pol21]. The body's response to stress involves various neuroendocrine reactions, primarily through the sympathetic nervous system (SNS) and the hypothalamic-pituitary-adrenocortical (HPA) axis. These pathways are crucial for secreting key stress markers like salivary alpha-amylase (sAA), cortisol, adrenaline, and noradrenaline [Ulr09] [Nat09]. While the effects of stress and its consequences are extensively studied, traditional methods for measuring neuroendocrine and electrophysiological markers are often labor-intensive for researchers. These methods are mostly invasive and costly, particularly those which rely on blood samples [Sla15]. Even noninvasive techniques tend to disrupt natural human behavior to some extent [Kaz79].

Video recordings represent a powerful method for capturing human behavior, providing a noninvasive and contactless way to observe facial expressions, voice characteristics, speech content, and head movements. These elements are vital indicators of human psychology and psychopathology, offering unique insights into individuals' emotional states [Dar72]. Advances in Computer Science have enabled the extraction of these markers from recordings - known as "Digital Biomarkers" [Ins17]. This technology bridges the gap between observational psychology and computational analysis, facilitating a deeper understanding of the nuances in human behavior.

Previous research has successfully linked distinct variations in facial expressions, speech, voice, and head movement features to mental and physical illnesses [Sch22a]. For example, digital biomarkers extracted from video-recorded interviews accurately predicted mental well-being in trauma survivors [Sch22b]. In stress measurement research, individual studies have explored changes in digital biomarkers such as speech behavior [Oes23], facial expressions [Gia17], and

movement [Ric22] due to acute stress. Efforts to integrate these biomarkers into a multimodal detection approach have shown promise, yet they have been limited by factors like small sample sizes [Aig18] and a narrow focus on male stress responses [Nor22b]. This highlights a significant gap in research: the need for a comprehensive study that combines digital biomarkers for a deeper understanding of acute stress reactions.

To gain a more comprehensive picture of acute stress responses, it is crucial to also consider internal physiological changes. The novel method of remote photoplethysmography (rPPG) can provide a promising solution by enabling the estimation of the heart rate (HR) remotely from facial videos, eliminating the need for physical contact [All07]. rPPG has its origins from photoplethysmography (PPG) which is a non-invasive optical technique used to detect blood volume changes in the microvascular bed of tissue, commonly used for measuring HR and oxygen saturation. rPPG extends PPG by using video recordings to capture these blood flow changes remotely, without the need for direct skin contact, allowing for HR monitoring from a distance using camera-based methods. Before deep learning, conventional rPPG methods like chrominance model (CHROM) [De 13] and plane-orthogonal-to-skin model (POS) [Wan17] dominated the rPPG landscape, using signal processing and machine learning to enhance HR detection accuracy by mitigating motion and lighting effects. However, conventional techniques still have difficulties handling datasets with challenging factors like varying lighting, high HR levels, and diverse skin tones, making them prone to noise degradation [Xia24].

Deep learning has since transformed remote HR measurement, introducing complex situation-adapted models like TSCAN (Convolutional Neural Network (CNN)-based) [Liu21a] and PhysFormer (transformer-based) [Yu22], which offer superior performance but require extensive data for training. However, despite their advancements, deep learning-based rPPG models face uncertainties in handling diverse skin tones, varying HRs, different lighting conditions, sudden movements, and have yet to be validated in scenarios involving speaking [Das21].

To obtain a more holistic understanding of acute stress reactions, the integration of digital biomarkers with rPPG-derived HR (rPPG-HR) offers a promising avenue for multimodal stress state detection. This approach could enable the assessment of cognitive and emotional states solely using markers extracted from upper body video recordings. Although one study on multimodal stress detection utilized a Trier Social Stress Test (TSST) dataset, it lacked a non-stressful control condition for comparison, leading to ambiguity in determining whether the classification was of stress or varying mental loads [Sab23]. This underscores the necessity for further research to validate HR metrics obtained from video for stress recognition purposes. Specifically, future models should aim to distinguish acute stress from different mental loads by incorporating a non-

stressful control condition and utilizing traditional markers to accurately control for physiological and subjective stress responses.

The goal of this master's thesis is, therefore, to explore the measurement of acute stress in a non-invasive and contactless manner using video analysis. It aims to develop a holistic approach to stress measurement by evaluating digital biomarkers derived from speech, facial expressions, and upper body movements. To achieve this, data from the EmpkinS collaborative research center's ongoing study, which involves healthy participants undergoing the TSST [Kir93] and its friendly control condition friendly TSST (f-TSST) [Wie13], was used. The TSST represents the gold standard for inducing acute psychosocial stress in controlled laboratory environments [Dic04]. In contrast, the modified version, known as the f-TSST, serves as a stress-free control condition that is as similar as possible to the TSST. In this study, the dataset will be further referred to as "Empkins-TSST".

The TSST and f-TSST were both video recorded, which enabled the extraction of digital biomarkers to analyze changes in facial expressions, speech, movement due to stress based on video and audio features from the Empkins-TSST. For this reason, a digital biomarker pipeline was developed.
To analyze the impact of acute stress on facial expressions, speech and movement, Machine Learning (ML) models were trained and evaluated to distinguish whether individuals were in exposed to acute stress or not. Furthermore, regression models were developed to examine the relationship between digital biomarkers and established biopsychological stress markers cortisol and sAA, assessed via saliva samples, as well as self-reported levels of stress, anxiety, and threat, assessed via questionnaires.

In this thesis, five different deep learning-based rPPG models DeepPhys, TSCAN, EfficientPhys, PhysNet, and PhysFormer were used to extract HR from facial videos. These models underwent cross-testing on two benchmark datasets and were also evaluated using the more dynamic Empkins-TSST dataset, which includes diverse HR levels, speech components, and motion variations. The findings from these deep learning models were then compared with the HR outputs obtained using five conventional rPPG techniques.

As a final step in the thesis, the effectiveness of stress state detection was assessed through a multimodal approach on the Empkins-TSST dataset. This involved augmenting the digital biomarkers with rPPG-HR, extracted from the best-performing rPPG model. These markers served as the basis for training ML models, enabling them to distinguish between stressed and non-stressed states.

To summarize, the thesis had the following four aims:

1. To predict acute stress versus a non-stressful control condition using video-based digital biomarkers and to analyze how facial expressions, body movements, and gaze behavior contribute to the classification output.

2. To examine the relationship between changes in facial expressions, speech, and movement, and traditional biological and psychological stress markers.

3. To determine whether the performance of conventional and deep learning (DL) rPPG methods decreased on more naturalistic datasets which address factors such as varying heart rate levels, speech sequences, and head movements.

4. To examine whether the combination of behavioral digital biomarkers with rPPG-derived HR (rPPG-HR) increases the stress prediction accuracy compared to behavioral digital biomarkers alone.

The thesis is organized as follows: Chapter 2 provides background information and reviews related work on digital biomarkers and rPPG. Chapter 3 delves into the Empkins-TSST dataset, which is utilized for analyzing acute stress reactions. This chapter further describes the digital biomarker pipeline and details the process of HR extraction from rPPG, along with an overview of the ML-based models employed for stress prediction. Chapter 4 presents the findings related to each aim of the thesis. Chapter 5 discusses the implications of using digital biomarkers and rPPG-HR, situating them within existing research. The thesis concludes with Chapter 6, which summarizes the main findings and suggests directions for future work in Chapter 7. Additional figures and tables are included in the Appendix for further reference.

Chapter 2

Background & Related Work

This chapter explores the basics of video-based digital biomarkers and the application of rPPG for HR extraction from videos, covering important background information and studies related to this thesis. It provides an overview of recent progress and the existing challenges in the area of acute stress detection and monitoring.

2.1 Digital Biomarker

This section explores how facial expressions, body movements, gaze behavior, and speech characteristics contribute to stress detection, providing background information and highlighting related work for each biomarker. Additionally, multidimensional stress detection discusses the combination of these video-based digital biomarker in a holistic model for identifying stress.

2.1.1 Facial Emotion Recognition

Background

Charles Darwin, in his work of 1872, postulated the concept of biologically "hard-wired" facial expressions of emotions [Dar72]. He proposed that these expressions serve as critical indicators, signaling and transmitting essential information about an individual's emotional and mental states. Building upon this foundational understanding, Ekman and et al. [Ekm78] later developed the Facial Action Coding System (FACS) - a methodology for systematically categorizing facial expressions [Ekm78]. The Facial Action Coding System (FACS) represents a comprehensive framework for analyzing human facial movements based on their appearance. Developed by

psychologists Ekman, Friesen, and Hager, it offers an invaluable tool for decoding emotional expressions by deconstructing complex facial gestures into individual Action Units (AUs) [Ekm78].

Emotions can be seen as the catalysts for a complex interplay of physiological responses and psychological experiences. They are reflected through an array of facial expressions that are often universally recognized [Etc92]. FACS encapsulates these expressions through a system of labeling various AUs, which correspond to discrete muscle movements. For instance, happiness is frequently associated with AU06 (Cheek Raiser) and AU12 (Lip Corner Puller), while anger is characterized by a combination of AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), and AU23 (Lip Tightener) [Ekm03]. The action units, their corresponding muscles, and associations are illustrated in Figure 2.1.

Table 2.1: Action Units related to emotions ("R" specifies only the right side, "A" stands for asymmetric.) [Ekm03]

Emotion	**Action Units**
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26
Fear	1+2+4+5+7+20+26
Anger	4+5+7+23
Disgust	9+15+17
Contempt	R12A+R14A

In the transactional model of stress, Lazarus and Folkman [Laz84] describe stress as a disparity between perceived challenges and available resources which result in an physiological overdrive [Laz84]. This heightened emotional state may activate specific AUs, such as AU1 (Inner Brow Raiser) and AU4, indicating psychological stress. Thus, facial expressions can offer valuable insights into an individual's emotional state.

Advancements in emotion recognition technologies have leveraged the specificity of FACS to interpret human emotions with increasing accuracy. By mapping AUs to corresponding emotional states, algorithms can now reliably predict emotions from facial cues, broadening the potential for applications in various fields, from psychology to human-computer interaction [Zha14].

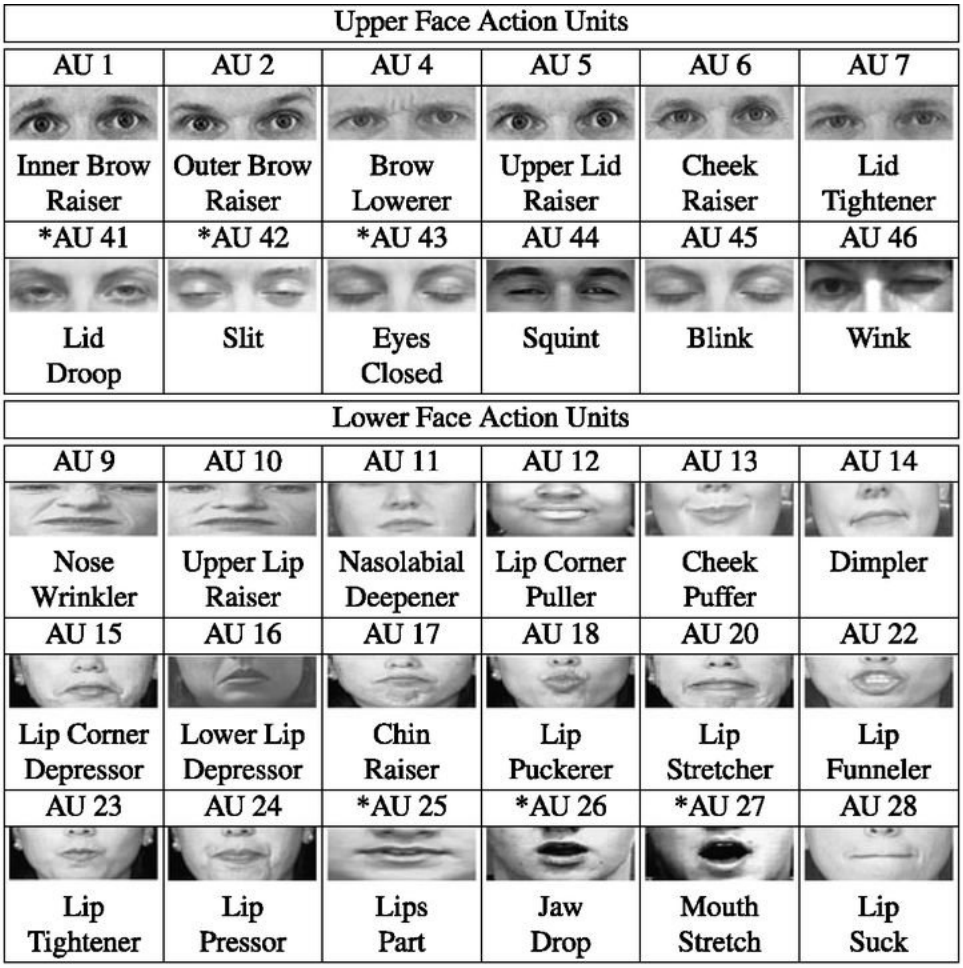

Figure 2.1: Actions units based on FACS and its corresponding facial muscles [Kun19].

Application

In a pioneering study, Lupis et al. [Lup14] conducted the TSST in 32 students and examined video-based facial expressions. The stress response was measured using the self-reported Positive and Negative Affect Schedule (PANAS) questionnaires and by assessing cortisol levels. Although no correlations were found between self-reported emotions and video-based facial expressions, the study revealed gender-related differences in the expressions of anger in response to psychosocial stress. These differences were observed using the FACS. Notably, men showed higher HR and cortisol stress responses associated with these anger expressions, a pattern that was not evident in women [Lup14].

Viegas et al. [Vie18] conducted a study consisting of seven different stress tasks and three neutral tasks to induce acute stress. Employing ML, they classified stressed and non-stressed/neutral phases based solely on AUs with an accuracy of 81.1%. Stressful tasks led to significantly increased AU intensities, resulting in a more "expressive" face. However, a limitation of this study was the absence of analyzing facial expressions in conjunction with subjective scores or physiological responses, raising questions about whether the classification was of rest vs. challenge (or arousal) rather than acute stress [Vie18].

In the recent pilot study "You Look Stressed", Blaseberg et al. [Bla23] employed a backward step-wise multiple linear regression approach to correlate subjective experience and objective stress responses from the body with facial activity due to exposure to the TSST. They observed no consistent pattern of facial activity across all stress markers. Their findings indicated that more frequent occurrences of the upper eyelid raiser (AU05) and upper lip raiser (AU10) were associated with higher stress-induced cortisol release, while more lip corner pulling (AU12) was linked to lower cortisol reactivity. The eyelid tightening (AU07) showed the strongest potential for detecting acute stress phases. Notably, the study revealed that women exhibited a greater intensity of stress-induced smiling, aligning with the "tend and befriend" hypothesis [Bla23]. This hypothesis suggests that tending involves nurturant activities designed to protect oneself and offspring, promoting safety and reducing distress, whereas befriending entails the creation and maintenance of social networks that may assist in this process [Tay00].

Although these studies collectively highlight the complex relationship between facial muscle activities and acute psychosocial stress responses, they also underscore existing challenges. These include the need for more comprehensive analyses combining facial expressions with subjective and physiological measures, and addressing gender-specific variations, thereby paving the way for more nuanced and accurate emotion recognition methodologies in stress-related contexts.

2.1.2 Bodily Movements & Stress

Background

Body movements and postures serve as non-verbal cues that offer a wealth of information regarding an individual's emotional and psychological state [Zit19] [Roe10] [Arn10]. For example, expansive gestures and relaxed postures typically signify positive emotions, whereas movements such as dropping the head or bringing hands to the face are associated with negative emotions like sadness [Wal86]. Human perception is finely tuned to interpret non-verbal cues from body postures, allowing for the subconscious differentiation between various emotional states. This capability plays a critical role in social interactions and the intuitive recognition of potential threats, reflecting a sophisticated cognitive function that is deeply embedded in human behavior [Ogr19].

The link between acute stress and changes in body movement is an area still in its infancy in stress research. Observations indicate that stress leads to noticeable decreases in movement amplitude and frequency — a potential behavioral response to perceived threats, signaling a shift in emotional state [Roe10]. However, in-depth studies are required to reliably utilize changes in body postures and movements under acute stress as markers for stress detection.

Current stress research focuses on quantifying these non-verbal indicators to objectively assess stress levels. The ability to measure and interpret the nuances of body movements offers promising avenues for healthcare applications. By translating these subconscious cues into quantifiable data, researchers aim to develop reliable biomarkers for the detection of acute stress.

Application

Until now, bodily movements have primarily been assessed using Motion Capture Suits (MoCap), which track movement through inertial measurement units (IMUs). Only a few studies have investigated differences in movement behaviors derived solely from video data.

In one of the pioneering studies aimed at detecting stress levels from movement data in video recordings, Giakoumis et al. [Gia12] induced stress in nineteen participants using the Stroop Color and Word Test (STROOP), during which they collected video, accelerometer, and biosignal (Electrocardiogram and Galvanic Skin Response) recordings. They successfully identified a set of activity-related behavioral features extracted from video recordings using Motion History Images [Bob01] that demonstrated a significant correlation with self-reported stress in their experimental evaluations [Gia12].

In a study by Roelofs et al. [Roe10], which included 50 participants, diminished body movements and a lowered HR were observed as a response to socially threatening cues, such as angry

faces. These physiological changes significantly correlated with subjective anxiety. Notably, the study exclusively involved women and used a stabilometric force platform to assess body sways [Roe10].

In a previous study using a protocol similar to this thesis, Richer et al. [Ric22] were able to classify individuals in a stressed condition from a stress-less control condition using MoCap data with an accuracy of 80%. The small sample size (18 participants) and a very unbalanced ratio of female to male participants may limit a good generalization of the results [Ric22].

Since only a few studies have analyzed bodily movements from video recordings, this highlights the need for further investigation into the potential of contactless movement assessments for acute stress detection.

2.1.3 Gaze Behavior & Stress

Background

Gaze behavior refers to the direction and focus of one's visual attention, indicating various cognitive and emotional states. It is a dynamic aspect of non-verbal communication, reflecting cognitive processes and situational engagement [Fri07]. In the context of acute stress, gaze behavior often becomes more restricted and focused. Studies have shown that under stress, individuals are likely to narrow their field of vision, concentrating their gaze more intensely on perceived sources of threat or concern [Bar07]. This shift in gaze pattern is thought to be part of the body's evolutionary response to danger, optimizing visual attention to better assess and respond to immediate threats.

Pupil diameter, regulated by the SNS, is another physiological indicator that varies with emotional and cognitive states [Bra08] [Kah66]. The dilation of the pupil, or mydriasis, occurs in response to increased SNS activity, often triggered by emotional arousal, cognitive load, or ambient light conditions [Wan18]. This response is hypothesized to enhance visual sensitivity and information processing to prepare the individual for a "fight or flight" decision.

Application

To date, only a few studies have explored the changes in gaze behavior due to acute stress. Herten et al. [Her17] employed eye-tracking glasses in a between-study design, comparing gaze behavior in the TSST with that in the f-TSST. Under stress conditions, participants demonstrated longer and more frequent fixations on central objects, along with reduced fixation times on committee faces, suggesting a tendency towards gaze avoidance of socially threatening stimuli [Her17].

Vatheuer et al. [Vat21] analyzed gaze behavior among 63 participants in a virtual reality adaptation of the TSST. They found a significant negative correlation between the duration of gaze on judges and cortisol production, indicating that individuals who spent less time looking at the judges exhibited a stronger cortisol response to acute stress. This suggests that gaze avoidance is linked to heightened stress reactivity. Moreover, an examination of gaze behavior across different phases of the TSST revealed higher gaze times on judges compared to surroundings during the speech task, while in the arithmetic task, this pattern was reversed [Vat21].

In their study, Guy et al. [Guy23] compared the speaking component of the TSST with a non-stressful control condition and observed a prolonged increase in pupil diameter following acute stress exposure through the TSST. However, the influence of changing light conditions on pupil diameter complicates its use as a proxy for acute stress responses. Gaze behavior was assessed before and after stress or control tasks, revealing that acute stress diminished visual exploration, evident from fewer saccades and a smaller scanned area. It didn't significantly affect attention to social features or image salience, challenging theories like Tend-and-Befriend, which suggest stress promotes social affiliative behaviors [Guy23].

Despite these insights, there remains a notable gap in research, as no studies to date have exclusively focused on analyzing gaze behavior under acute stress through video recordings alone, a method potentially more applicable and versatile in various research contexts than eye-tracking glasses or virtual reality setups. Given the complexity of extracting gaze patterns from videos, there is a pressing need for more research in this area.

2.1.4 Speech Characteristics & Stress

Background

The human voice serves not only as a primary means of communication but also as a potent indicator of an individual's psychological and physiological states [Bel04]. Voice production is a complex physiological process initiated by the air pressure system, particularly the lungs, which generate the necessary airstream pressure. This stream of air is then intricately shaped by the phonatory system within the larynx, where the vocal folds engage in vibratory motion, a phenomenon known as phonation. Beyond the larynx, the articulatory system, including the nasal and oral cavities along with the pharynx, adds resonance, while the coordinated action of the tongue, palate, and lips articulates the phonated sound into distinct speech elements [Hon08]. A general scheme of the voice production apparatus is shown in Figure 2.2.

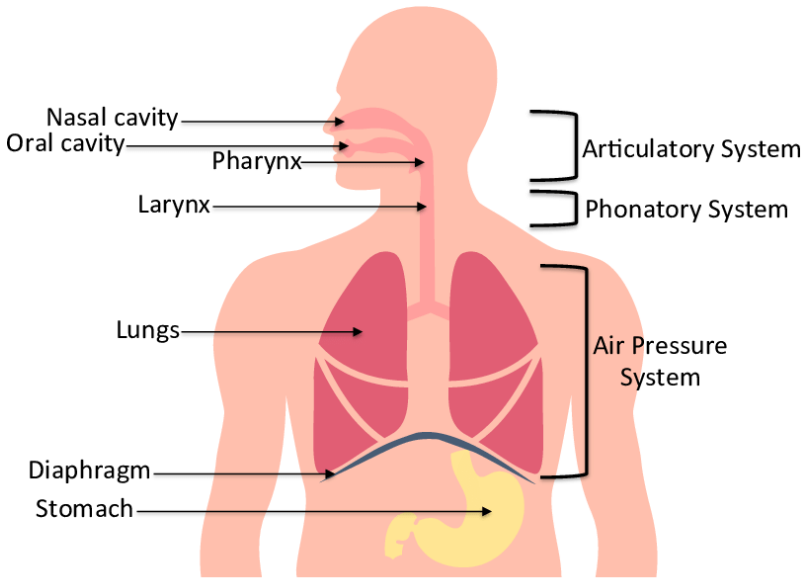

Figure 2.2: General scheme of the voice production apparatus [And17].

The qualities of voice – pitch, volume, and timbre – are the result of this complex interplay. Under stress, these qualities exhibit measurable changes. For stress detection, the voice signal is analyzed in the time and frequency domain. Time-domain analysis captures features such as jitter, which measures the frequency variation for specific time periods, and shimmer, which reflects amplitude variations. In the frequency domain, spectral features are examined, including formant frequencies that are resonant frequencies of the vocal tract and are often affected by stress [De 06].

Voice analysis for stress detection offers a non-invasive method to monitor mental health and aid in stress management. Ongoing advancements in machine learning promise to enhance the sensitivity and specificity of stress detection tools, which could lead to significant improvements in digital health applications.

Application

The study of speech in relation to psychosocial stress is still evolving, with current research efforts primarily directed towards the identification and refinement of fundamental speech parameters. A common observation regarding stress-induced changes in speech is an increase in the fundamental frequency (F0), as noted by Kirchbühel et al [Kir11]. Additionally, formants have emerged as promising speech features for distinguishing stress. In particular, shifts in the first formant (F1) and second formant (F2) have been identified by Van Puyvelde et al. as reliable indicators of psychological stress [Van18].

In a pioneering study, Baird et al. [Bai19] assessed the ability of speech-based features to predict sequential cortisol measurements. They observed a moderate correlation between speech characteristics and cortisol levels during the TSST [Bai19]. In subsequent research involving 100 participants, they extended their analysis to forecast physiological parameters like HR and respiration, employing the same set of features in conjunction with a DL-based framework [Bai21].

In the most recent study, Oesten et al. [Oes23] evaluated the efficacy of vocal acoustics for stress detection in a within-subjects design (TSST vs. f-TSST). Their research revealed significant alterations in acoustic features under stress. A stepwise backward multiple linear regression model explained 58.8% of the variance of the maximum cortisol increase. Additionally, their classification experiments were able to distinguish between stressed and non-stressed states with an accuracy of 80.0 ± 12.7%. However, the small sample size of 21 participants, predominantly female (85.7%), may limit the generalizability of these findings [Oes23].

Due to the controlled nature of speech content in the TSST, only a few studies have analyzed speech and language output during its speech component. Buchanan et al. [Buc14] noted that stress resulted in increased pause times during speech; however, they also observed that language productivity was actually higher under stress compared to non-stressful speaking tasks. In contrast, participants exhibited more frequent pauses during the stressful TSST, a trend that was particularly pronounced among those who exhibited larger cortisol and HR responses to the stressor. This study highlights the intricate interplay among stress, speech, and language, and supports the anecdotal evidence of stress-impaired speech production abilities [Buc14].

Given the considerable variability in speech features and the pronounced differences observed between genders, it is imperative to analyze speech within a larger, more diverse sample and through a within-study design to enhance the robustness and generalizability of findings in the context of acute stress detection.

2.1.5 Multidimensional Stress Detection in Videos

Background

Facial expressions, body movements, gaze behavior, and speech characteristics each offer unique insights into an individual's emotional state. The human brain has the remarkable ability to integrate these diverse signals into a coherent perception of others' emotions, much like an orchestra combines different instruments to create a harmonious composition [Foa86]. This holistic processing allows for a more accurate assessment of emotional states, including acute

stress, by recognizing the nuanced interplay among tightened facial expressions, constrained body movements, focused gaze, and speech alterations [Van07].

Currently, researchers are exploring ways to combine digital biomarkers, including facial expressions, gaze behavior, body posture, and vocal characteristics from videos, into a single multidimensional model. This holistic approach aims to gain a better understanding of how these indicators interplay in conveying stress and other emotional states.

Application

In the study by Pisanski et al. [Pis18], 80 adults underwent the TSST, with voice, polygraph, and hormone measures recorded. The findings revealed that stress led to an increase in voice pitch, along with a decrease in hand movements and skin temperature, with striking similarities between men and women. However, the relationship between cortisol, skin temperature, and voice pitch changes was weak, indicating inconsistencies in stress responses across genders. This highlights the urgent need for further research using multimodal stress measures to explore the factors behind these varied responses, including gender and individual stress reaction differences [Pis18].

In the study by Aigrain et al. [Aig18], multimodal stress detection was explored through a combination of physiological data, subjective markers, and objective behavioral markers such as body movement and facial expressivity, all derived from video recordings. The study focused on a socially evaluated mental arithmetic test with a sample size of 25 participants. It achieved a notable F1 score of 0.85 in predicting stressed versus non-stressed states using these multimodal features. The findings underscored that integrating data from blood volume pulse, HR, movement behavior, and facial expressions results in a more robust prediction of stress. However, the study was limited by its small participant pool of only 25 individuals and the absence of a control condition [Aig18].

Zhang et al. [Zha20] conducted a study on stress detection using a deep learning-based model that analyzed facial expressions and action motions from video recordings. The model achieved an accuracy of 85.42% in identifying stress states. In their experiment, stress was induced by having participants watch a scientific program and then answer ten questions about each video. For the non-stressful task, participants watched neutral videos. This method aimed to simulate stress-inducing conditions, though the actual stress response was not directly quantified. The study included a sample size of 122 participants [Zha20].

Norden et al. [Nor22b] explored various stress dimensions, examining how the choice of stress labels might influence predictive performance in ML pipelines. They trained three standard ML models to predict different stress labels using either voice or facial cues, with only male participants (N=40) involved. ML models showed that voice-based predictions correlated strongly with

panel-annotated stress levels (ρs = .54), outperforming other models (ρs = .30). Face-based models were positivly related to ground truth values for face-based (ρs = .24) but not for voice-based models. However, predictions for video-annotated stress and endocrinological stress levels were not successful in both settings. They suggest that future work should incorporate other potentially stress-relevant cues, such as head-pose, posture, eye-gaze, and linguistic features, from the videos into the prediction model [Nor22b].

So far, the exploration of digital biomarkers for contactless stress detection has yielded promising results. However, the presence of challenges such as limited sample sizes and the absence of control conditions underscores a critical need for further research. This necessitates the incorporation of both subjective and objective stress assessment markers to accurately predict stress responses, including physiological markers like HR. For a more holistic understanding, investigating the efficacy of rPPG methods in accurately predicting physiological changes presents a promising avenue for advancing contactless stress detection technologies.

2.2 rPPG

In recent years, rPPG has become increasingly popular for contactless HR extraction from facial videos. The subsequent sections will delve into the origins and initial models rPPG, followed by an exploration of its current validation and application in acute stress research.

2.2.1 Background

From PPG to rPPG

Building on the established PPG method, remote photoplethysmography (rPPG) is a pioneering, non-contact method for HR measurement. PPG technology employs a light source and a photodetector to monitor changes in blood volume under the skin, based on the Beer–Lambert law which associates light absorption with hemoglobin concentration [McD15]. This principle supports non-invasive devices such as pulse oximeters and fitness watches. Yet, traditional PPG devices face limitations, including unsuitability for sensitive individuals, discomfort, and potential inaccuracies due to environmental variables [Al-17].

To address these issues, non-contact rPPG methods have been developed, utilizing camera technology to detect subtle skin color variations and derive the PPG signal. Unlike traditional PPG that requires physical contact, rPPG methods analyze video recordings of the subject's

face [McD15].

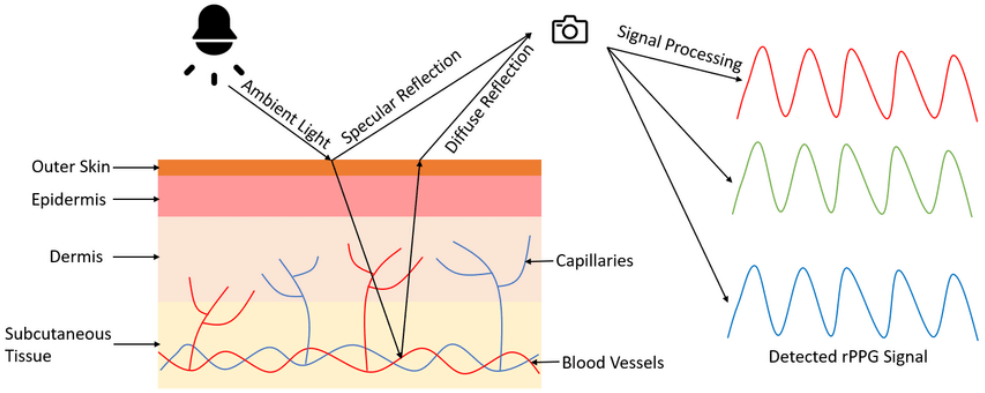

Figure 2.3: Diagram of rPPG signal generation, showing a camera capturing specular (non-informative) and diffuse (blood volume-related) skin reflections under environmental light, enabling rPPG signal extraction.

The operational basis of rPPG is akin to conventional PPG, with the distinction being in signal acquisition. rPPG processes video to observe blood flow changes with each heartbeat. As illustrated in Figure 2.3, the Dynamic Region Model (DRM) [De 13] elucidates rPPG's functionality, distinguishing between specular and diffuse reflections from the skin. While specular reflection provides no physiological information, diffuse reflection, emanating from blood vessels, is rich in physiological data [Ver08]. rPPG focuses on extracting this meaningful diffuse reflection.

Advantages of rPPG include eliminating the need for contact-based devices, thus mitigating their associated drawbacks, and its suitability for continuous, long-term monitoring across various patient populations. However, the reliability of HR extraction via rPPG methods remains in question, highlighting the urgent need for additional validation before it can be considered a viable alternative to traditional HR monitoring techniques [Xia24].

Conventional Models

Before the emergence of DL, conventional methods dominated the rPPG landscape. The feasibility of rPPG was first demonstrated in 2008 by Verkruysse et al. [Ver08], marking a pivotal moment that paved the development of several conventional techniques [Ver08]. Conventional rPPG methods primarily leverage algorithms to reduce motion artifacts and noise in facial videos for improved signal extraction. Techniques like Independent Component Analysis (ICA) [Poh10] and Principal Component Analysis (PCA) [Lew11] struggle with artifacts that share frequencies with the normal HR range.

Model-based approaches, notably the CHROM [De 13] and the POS [Wan17], utilize the blood volume pulse changes to distinguish between pulse-induced and motion-induced color changes in RGB videos. Another method called blood volume pulse (BVP) [De 14], same name as the underlying blood volume signal itself, differentiates pulse-induced color shifts from motion artifacts in the RGB spectrum by leveraging characteristic blood volume changes. Another method called 2SR is an innovative algorithm enhancing rPPG accuracy by estimating skin-pixel subspace rotation, bypassing the need for skin-tone or pulse assumptions. It surpasses conventional methods like ICA and CHROM, detecting pulse rates accurately under various conditions [Wan16].

A study comparing POS, ICA, PCA, CHROM, and BVP across various conditions highlighted the effectiveness of POS in both stationary and motion contexts, despite challenges in differentiating similar amplitude signals. In particular, POS has shown superior performance in handling high noise scenarios like fitness challenges due to its physiological basis [Wan17].

Conventional methods still face difficulties with datasets under challenging conditions, including varying lighting, elevated HRs, and diverse skin tones, making them susceptible to noise degradation [Xia24]. However, emerging DL-based rPPG methods offer promising approaches to overcome these challenges.

Deep Learning Models

deep learning (DL) advancements have revolutionized remote HR measurement by introducing models capable of handling complex situations. These predominantly supervised methods require extensive training data but deliver superior performance. Since the second part of the thesis primarily focuses on validating rPPG methods, only publicly accessible pre-trained models of supervised methods will be introduced and discussed.

Inspired by the success of 2D CNNs in computer vision, Špetlík et al. [Spe18] introduced the first rPPG DL method, HR-CNNs [Spe18], designed for remote HR measurement using a two-stage CNN architecture. Validated on the PURE dataset [Str14], HR-CNN significantly outperformed traditional methods with a Mean absolute error (MAE) of 3.14 beats per minute (bpm) and demonstrated robustness against video compression.

Following HR-CNN, Chen et al. proposed DeepPhys [Che18b], which builds on the success of 2D CNN with a focus on motion robustness, incorporating an attention mechanism for improved performance. However, its limitation in capturing temporal information was addressed by Liu et al. with MTTS-CAN [Liu21a], which integrated the Temporal Shift Module (TSM) for better

temporal analysis, demonstrating notable improvements on the UBFC-rPPG dataset [Bob19].

Yu et al. then introduced PhysNet [Yu19], a 3D CNN method that directly extracts rPPG signals from raw RGB frames, showcasing the advantages of 3D over 2D CNNs in learning spatial and temporal features. PhysNet's effectiveness was proven through its performance on the OBF dataset and its potential application in emotion recognition. Further, Yu et al. also developed Physformer [Yu22], employing a time-difference transformer for advanced spatiotemporal analysis. This method stands out for its precise signal estimation and robust performance across various datasets, highlighting its state-of-the-art capabilities.

Despite these advancements, the effectiveness of rPPG highly depends on the training data and remains sensitive to illumination, movement, and skin color variations. Moreover, Zhan et al. highlighted limitations in the robustness of DL models and suggested exploring hybrid CNN methods, integrating PPG-related knowledge for improved outcomes, indicating a nuanced path forward in rPPG research [Zha20].

2.2.2 Application

Validation of rPPG Models

Currently, accessing code for various learning-based methods is challenging. This is indicated by the fact that only few studies actually validate developed state-of-the-art rPPG methods on new and more naturalistic datasets [Xia24].

Liu et al. [Liu23] compared six convolutional methods (GREEN [Ver08], ICA, CHROM, LGI, PBV, POS) with five DL-based methods (TS-CAN, PhysNet, PhysFormer, DeepPhys and EfficientPhys) on the benchmark datasets PURE and UBFC-rPPG. PhysNet outperformed on UBFC-rPPG with a MAE of 0.98 bpm, followed by Physformer (MAE: 1.44 bpm), both trained on PURE. Regarding the PURE dataset, POS performed the best with a MAE of 3.67 bpm closely followed by TS-CAN trained on UBFC-rPPG (MAE: 3.69 bpm). The above validation was part of the development of the rPPG-Toolbox which contains unsupervised and supervised rPPG models with support for public benchmark datasets, data augmentation, and systematic evaluation to further facilitate the validation of state-of-the-art rPPG models [Liu23].

In the review paper by Ni et al. [Ni21], the benchmark UBFC-rPPG dataset was used to compare the performance of four DL methods for HR measurement, namely rPPGNet, 3D-CNN, PhysNet and Meta-rPPG. The results showed that PhysNet generated the best HR measurement outcome among these methods, with a MAE of 2.57 bpm and a MSE of 7.56 bpm [Ni21].

Yang et al. [Yan22] evaluated the performance of three DL-based methods (DeepPhys, rPPGNet, and Physnet) to that of four traditional methods (CHROM, GREEN, ICA, and POS) using two public datasets: 1) UBFC-rPPG; 2) the BH-rPPG. The experimental results demonstrate that traditional methods are more resistant to fluctuating illuminations. They found that the PhysNet achieved the lowest MAE among DL-based method under medium illumination, whereas the CHROM achieves 1.04 bpm, outperforming the PhysNet by 80% [Yan22].

Research indicates that testing state-of-the-art rPPG models on more real-world datasets is still insufficient to validate the reliability of the estimated HR. This validation is crucial for applications in medical contexts or stress state recognition.

Stress State Recognition with rPPG

Combining remotely assessed physiological states through rPPG with digital biomarkers unveils the potential of exclusively contactless methods for identifying emotional and cognitive stress levels solely through video recordings.

McDuff et al. [McD14] used ICA-based rPPG to predict stress states from HR, breathing rate, and heart rate variability (HRV), which aligned well with contact-based PPG measurements. Despite no marked HR differences between relaxed and stressed states in their ten-participant study, the model predicted stress with 85% accuracy using a person-independent classifier. The small sample size and lack of cross-validation suggest potential overfitting, and findings of only HRV differences raise questions about classifying mental rather than stress states [McD14].

In a study by Iuchi et al. [Iuc20], seven participants performed mental arithmetic under varying complexities with fixed head positions, resulting in 78 videos. They developed a new method called HMS+2SR which first converts pixel values to hemoglobin components named HMS [Fuk17], and then extracts pulse waves using 2SR. Seven different rPPG methods were tested, finding 2SR most accurate (Mean absolute percentage error (MAPE) nearly 80%). Using k-NN, 2SR and HMS+2SR showed 70% accuracy in stress classification, with ICA also above 70%. The study did not compare HR variations or State-Trait Anxiety Inventory (STAI) scores across stress levels and was limited by its small sample and restricted head movements [Iuc20].

Morales-Fajardo et al. [Mor22] extracted HR from webcam videos using rPPG of fifty-six undergraduates to assess academic anxiety. Together with demographic data, they compared normal classes to exams with STAI questionnaires. Achieving 96% accuracy with kNN, J48, and Random Forest classifiers, the project's HR data lacked validation against a ground truth, casting uncertainty on its role in predicting anxiety alongside factors such as gender, school type, and

extracurricular activities [Mor22].

Benezeth et al. developed an rPPG algorithm for HRV estimation, demonstrating a strong correlation with emotional states [Ben19]. In a follow-up, Sabour et al. recorded participants during the TSST, using POS to extract HRs from videos. The study noted the lowest MAE during rest (3.55 bpm), then the math task (5.99 bpm), and the speech part (9.26 bpm), achieving 85.48% accuracy in stress detection with rPPG and electrodermal activity (EDA) features. This marked the first dataset to include speech and more natural movements in rPPG extraction [Sab23].

The studies highlight the need for further validation of HR extracted from videos, especially for stress recognition models to distinguish actual stress from different mental loads. This thesis aims to validate both DL and conventional rPPG methods in real-world scenarios, taking into account HRV, speech sequences, and head movements. Additionally, it will examine whether HR, combined with other digital biomarkers, can effectively identify stress.

Chapter 3

Methods

3.1 EmpkinS-TSST Dataset

To evaluate the impact of acute stress on the body using digital biomarkers, data from an ongoing video-recorded research study at the *Machine Learning and Data Analytics Lab* by the EmpkinS collaborative research center [Emp24] were utilized. In this study, participants underwent the TSST and f-TSST on two consecutive days. To ensure a balanced dataset, the condition order was randomized.

3.1.1 Study Population

For this thesis, data from 44 participants (25 female and 19 male) were analyzed. An overview of the gender distribution and the order of conditions is presented in Table 3.1, while demographic and anthropometric data for the participants are detailed in Table 3.2.

Table 3.1: Gender distribution of the first part of the EmpkinS TSST study dataset.

Condition order	Female	Male	Total
f-TSST-first	13	10	23
TSST-first	12	9	21
Total	25	19	44

Participants were recruited through electronic flyers on social media and mailing lists. They filled out a screening questionnaire to determine eligibility, with exclusion criteria including being outside the 18-50 age range, non-German native speakers, Body Mass Index (BMI) outside 18-30, any physical/mental illness, medication use, smoking, drug use, adiposity, or previous similar stress

Table 3.2: Demographic and anthropometric distribution of the first part of the EmpkinS TSST study data (mean ±std).

	Age [years]	**Height [cm]**	**Weight [kg]**	**BMI [kg/m^2]**
Female	22.65 ± 2.68	168.70 ± 5.94	59.90 ± 8.24	20.95 ± 1.80
Male	22.75 ± 3.32	181.63 ± 6.47	75.31 ± 8.58	22.84 ± 2.46
All	22.70 ± 2.90	174.32 ± 8.82	66.84 ± 11.20	21.86 ± 2.29

test participation. As compensation, they could choose between receiving 50€ or, for psychology students, 5 "Versuchspersonenstunden".

3.1.2 Acute Stress Induction

In the ongoing EmpkinS study, which employs a within-subjects design, the TSST is utilized for stress induction—recognized as the gold standard for eliciting acute psychosocial stress in laboratory settings [Dic04]. The f-TSST serves as a control condition, specifically designed not to activate the HPA axis or increase negative affect [Wie13]. Differences between the f-TSST and TSST protocols within this design are illustrated in Figure 3.1.

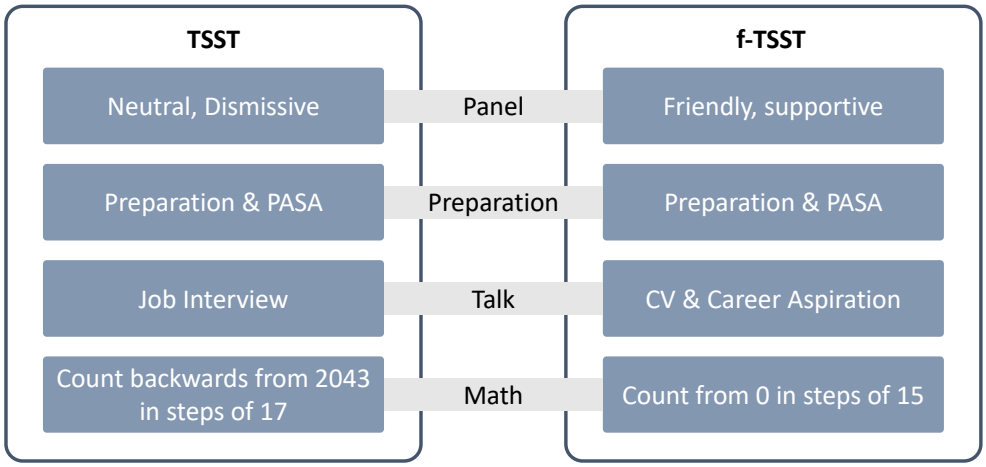

Figure 3.1: Differences between TSST and f-TSST.

TSST

The TSST involves participants performing in front of a neutral two-person panel, consisting of a male and a female experimenter in white lab coats. The active member of the panel, always of the opposite gender to the participant, is seated on the right site of the panel. The protocol encompasses

three phases, each lasting 5 minutes: preparation, talk, and math. During the preparation phase, participants are briefed that the panel will judge their suitability for their dream job based on personality traits. They then prepare notes for an interview-style conversation and complete the Primary Appraisal Secondary Appraisal (PASA) questionnaire [Gaa09] to assess cognitive appraisals under stress [Car16]. After the preparation, participants begin their interview, speaking continuously, with the panel intervening only if the participant strays off topic or remains silent for more than 20 seconds. In the final math phase, participants perform a mental arithmetic task, counting backwards from 2043 in steps of 17, with instructions to restart at any mistake.

f-TSST

Like the TSST, the f-TSST is conducted in front of a two-person panel but aims to minimize stress while maintaining a comparable setting. The panel does not wear lab coats and adopts a friendly, supportive demeanor. During the preparation phase, the opposite-gender panel member exits the room. The talk part is designed as a conversation about the participant's CV and career aspirations, rather than a stressful job interview. Since the original f-TSST protocol does not include a math component, the math section from the placebo-TSST [Het09], a low-stress alternative to the TSST, is utilized. Here, participants count in steps of 15 starting from 0, with errors pointed out in a supportive manner and instructions to continue from the last correct number.

3.1.3 Study Procedure

In the following, a brief overview of the EmpkinS-TSST protocol is given.

Pre-Test

Upon arrival at the lab, participants entered the preparation room to provide consent to participate in the study before undergoing initial assessments: the first saliva (S0) was collected. To balance glucose levels in the blood, 200 ml of grape juice or sugar water was provided [Zän20]. Female participants also provided a sample for progesterone level analysis to account for menstrual cycle effects on cortisol response [Ham20].

Participants' body composition was measured using a scale for body weight, fat, and muscle percentage. Electrocardiogram (ECG) data were captured via a ECG sensor node (Portabiles GmbH, Erlangen, Germany), attached to a chest strap, recording a 1-channel ECG according to Lead I of Einthoven's Triangle with a sampling frequency of 256 Hz.

(f-)TSST

Approximately 40 minutes after arrival, participants were brought to another room to undergo the (f-)TSST, detailed in Section 3.1.2. During the (f-)TSST, facial expressions and body movements were recorded using an red-green-blue (RGB) camera (Sony SRG-300H) and an RGB-D camera (Microsoft Azure Kinect), while a smartphone app logged test phases. Video and ECG data synchronization utilized Inertial Measurement Unit (IMU) attached to both the camera and ECG, recorded on the same system for matching timestamps. A peak was generated in each signal for alignment. An IMU on a clapperboard, captured in the video, allowed manual synchronization by matching the clap moment and the IMU signal peak.

Post-Test

Directly after the (f-)TSST, participants returned to the preparation room for a saliva sample (S2), followed by self-report questionnaires, including the PANAS [Wat88]. A complete overview of the questionnaires is provided in Table 3.3. Additional saliva samples were collected at post-test intervals (S3 - 25min, S4 - 35min, S5 - 40min, S6 - 60min, S7 - 75min). At the end of the testing day, participants were either reminded of their next session (Day 1) or debriefed and asked to sign a non-disclosure agreement about the study setup (Day 2).

3.1.4 Stress Response Measures

The stress assessment comprises of objective (HR, cortisol, sAA) and subjective measures (self-report). In the following each one will be presented and how they were further processed.

ECG

The acquired ECG data were processed to extract R-peak-to-R-peak (RR) intervals, following an initial filtering step and the application of a QRS detection algorithm provided by the Neurokit2 library [Mak21]. To minimize artifacts in the RR intervals, methods from prior studies were employed, such as those described in [Hap21]. From these refined RR intervals, HR was calculated to assess the physiological responses of participants. Due to the absence of a true ECG baseline measurement in the study design, it was not possible to calculate the percentage increase in HR relative to a baseline (ΔHR). Furthermore, HRV metrics, including $RMSSD$, $pNN50$, and $SD1/SD2$, were determined following the guidelines set by the HRV Task Force [Mal96], providing insight into sympathetic nervous system activation.

Endocrinological Data

Using Salivettes (Sarstedt AG & Co. KG, Numbrecht, Germany) for collection, the eight samples remained at room temperature until session completion and were then frozen at $-18°$C. Later, cortisol concentrations were extracted from the saliva samples using established procedures detailed in prior research [Ric21a].

To further quantify the stress response, four cortisol parameters were derived from raw cortisol values, following the methodology outlined by Pruessner et al. for AUC_g and AUC_i [Pru03]:

c_{max}	Maximum cortisol increase
m_{S1S4}	Slope from S1 to S4
AUC_g	Area under the curve with respect to ground
AUC_i	Area under the curve with respect to increase

The maximum cortisol increase c_{max} is determined by the difference between the highest cortisol level after the (f-)TSST and the initial level (Equation 3.1). The equation is defined as:

$$\Delta c_{\text{max}} = \max(S_i) - S_1, \quad \forall i \in [2, 7] \tag{3.1}$$

The slope between the first and fourth samples, m_{S1S4}, is calculated using the formula:

$$m_{S1S4} = \frac{S_4 - S_1}{t_4 - t_1} \tag{3.2}$$

The total cortisol released in response to the (f-)TSST, denoted as AUC_g, is calculated with the trapezoidal rule (Equation 3.4), where S_i represents the cortisol level at time t_i, and Δt_i is the time interval between consecutive samples in minutes.

$$AUC_g = \sum_{i=1}^{6} \frac{(S_{i+1} + S_i) \cdot \Delta t_i}{2} \tag{3.3}$$

Lastly, AUC_i measures the area under the curve with respect to the initial increase, calculated as:

$$AUC_i = (\sum_{i=1}^{6} \frac{(S_{i+1} + S_i) \cdot \Delta t_i}{2}) - (t \cdot S_1) \tag{3.4}$$

Self-Report Data

During the study, participants completed psychological questionnaires at three stages: screening, pre-test, and post-test. Initially, eligible participants - as defined in Section 3.1.1 - completed a set of questionnaires during screening. They then completed another set before and immediately after each (f-)TSST intervention. A comprehensive list of these questionnaires is presented in Table 3.3 This thesis will specifically analyze the PANAS questionnaire.

To evaluate the effects of the (f-)TSST on various psychological state variables, the differences in questionnaire scores before and after the (f-)TSST were calculated. Specifically, for PANAS, two distinct scores were analyzed: positive affect (PANAS-PositiveAffect) and negative affect (PANAS-NegativeAffect).

Table 3.3: Overview of psychological questionnaires used in the study.

Set	**Screening**	**Pre**	**Post**
Questionnaires	ADS-L	STADI State	STADI State
	STADI Trait	PANAS	PANAS
	Brief-Cope	SSSQ	SSSQ
	PSS		VAS
	BFIK		SSSGS
	RSE		
	SCS-D		
	RSQ		
	BES		
	SOC		
	TSGS		

3.2 Digital Biomarker

In this thesis, a pipeline was developed for extracting digital biomarkers from videos. This pipeline is comprised of two main components: pre-processing and feature extraction, as highlighted in Figure 3.2. Due to the poor audio quality of the TSST dataset (outlined in Section 3.1), the focus of this thesis is limited to video-based digital biomarkers, such as facial expressivity, movement patterns, and gaze behavior. The subsequent sections will detail the methods used for extracting these features from videos. Furthermore, it is shown how these digital biomarkers were analyzed to assess the impact of stress and their relationship with traditional biological and psychological stress markers.

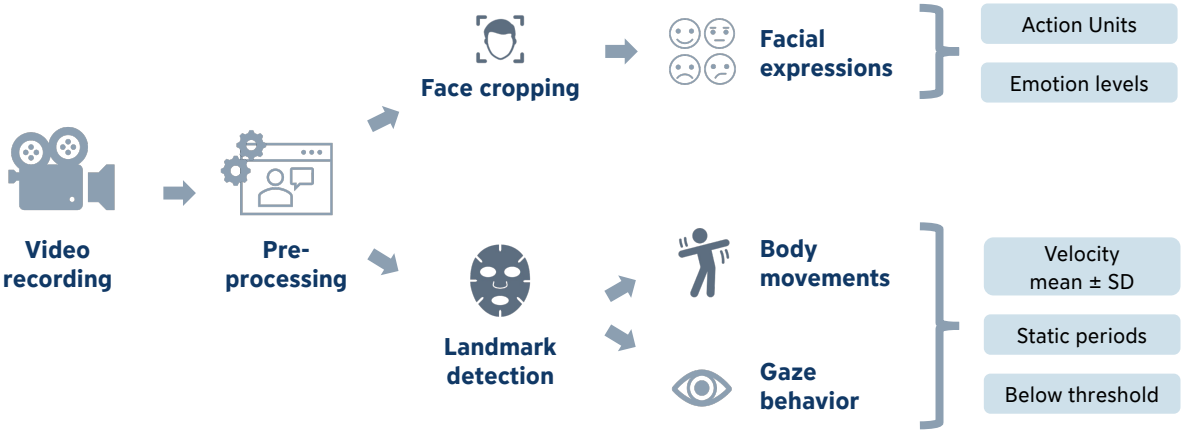

Figure 3.2: Digital Biomarker Pipeline: Video processing to extract facial expressions, movement patterns and gaze behavior.

3.2.1 Pre-processing

Video Synchronization

As outlined in the study design in Section 3.1.3, IMUs were employed to synchronize the video recordings with the ECG signal. To facilitate this synchronization, a peak was generated in each signal for alignment purposes: one peak through the act of clapping a clapperboard for the video, and another peak generated within the ECG signal itself. For ECG synchronization, the peak in the ECG signal was synchronized with the corresponding peak in the IMU data. If a peak for ECG synchronization was not detected, the original timestamp from the ECG IMU was used as a fallback. The peak created by the clapperboard was then employed to align the timestamp from the video recording with a manually detected timepoint within the video. In instances where the

peak for camera alignment was absent, synchronization of the video recording was achieved by mapping the timestamp from the start of the talk phase to the corresponding timepoint in the video recording.

Cut Video Phases

Following the synchronization of the video recording, videos were segmented into the distinct phases of the (f-)TSST for detailed analysis.

3.2.2 Features

To analyze the impact of acute stress on facial expressions, movement patterns, and gaze behavior, the following features were extracted. A detailed overview can be found in the Appendix B.

Facial Expression Recognition

Facial expressions and AUs were analyzed from video data using the open-source Pyfeat package [Che21], which supports pre-trained models for face detection, landmark identification, emotion recognition, and AU detection. To optimize computation, only one frame per second was processed, reducing the original frame rate from 29 fps to 1 fps.

Initially, video data requires pre-processing for both emotion levels and AUs values. This step involves detecting faces and identifying landmarks, which are crucial for classifying emotions and AUs values. PyFeat offers a comprehensive pipeline that integrates both pre-processing and facial expression classification into one system. For face detection, it employs RetinaFace - a single-stage method for dense face localization developed by Den et al. [Den19]. The detected faces are then used to extract facial landmarks with MobileFaceNet [Che18a], an efficient CNN designed for real-time face verification with high accuracy. Finally, the facial landmarks serve as an input for subsequent classification of emotion levels and action units.

Emotions
The residual masking network, ResMaskNet [Pha21], an end-to-end CNN model that integrates deep residual networks with masking blocks, was utilized for emotion detection. By focusing on local regions, ResMaskNet refines its predictions and maintains performance in deeper layers. It assigns a probability to each frame for seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral.

Action Units (AUs)
Pyfeat employs the XGBoost Classifier to identify 20 distinct AUs. Leveraging the architecture of OpenFace [Bal18], the model processes and compresses Histogram of Oriented Gradients (HOG) features from facial landmarks with a convex hull algorithm and PCA, predicting the AU through ensemble-based shallow learning techniques. Each AU receives an activation value on a scale from 0 to 1 for every frame.

Descriptive statistics, including mean, standard deviation, minimum, and maximum values, were calculated for both emotion levels and AUs values over the different phases. The comprehensive list of features is shown in Table 3.4.

Table 3.4: Overview of emotion values and AUs levels.

Type	Feature	Metric
emotions	anger, disgust, fear, neutral, happiness, sadness, surprise,	mean std
action units	AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU11, AU12, AU14, AU15, AU17, AU20, AU23, AU24, AU25, AU26, AU28, AU43	min max

Movement Patterns

To extract movement features from videos, facial landmarks need to be detected. Each landmark is characterized as one pixel value in the x-y plane. In this thesis the package MediaPipe [Lug19] from Google was used to extract facial and upper body landmarks. MediaPipe Face Mesh provides a robust framework for real-time estimation of 468 3D face landmarks using machine learning, requiring only a single camera input without additional depth sensors. It features an efficient pipeline that includes a face detector as well and a 3D face landmark model, designed to deliver high accuracy in facial surface mapping by utilizing lightweight architectures and GPU acceleration. The 468 facial landmarks are shown in Figure in detail 3.3.

To ensure the observation of pure head movements, while excluding muscle activity related to facial expressions, this study designated facial landmarks around the nose area for the quantification of head movements. The mean value of these 10 landmarks served as the basis for further analysis. Following previous work, movement features were divided into two types: generic and expert [Ric22]. Generic features encompass mean distance, range of motion, mean velocity, and standard deviation of movements, all of which do not require domain-specific knowledge. Further, the visibility of both hands and elbows was calculated by counting their appearances

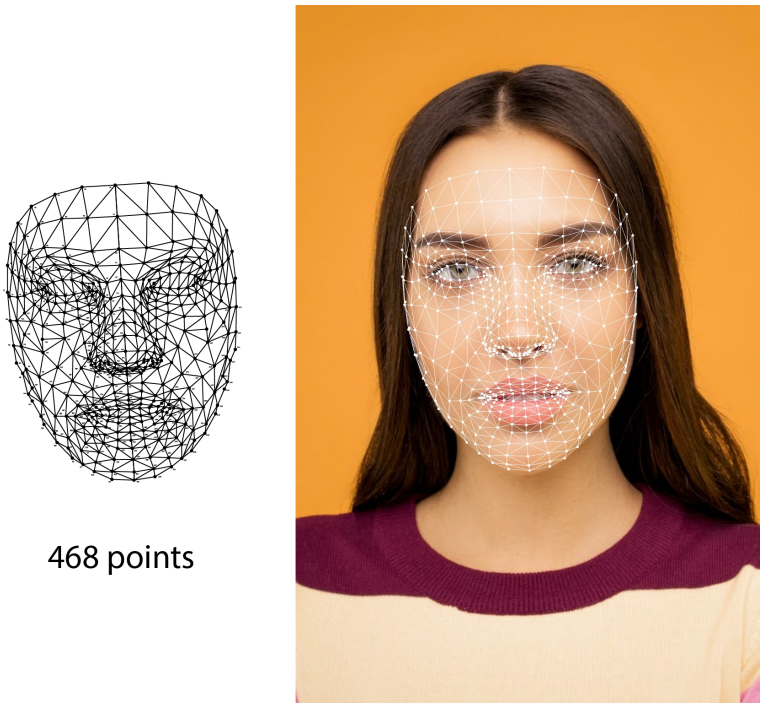

Figure 3.3: Mediapipe 3D FaceMesh with its corresponding facial landmarks, as depicted by Google Mediapipe.

in the video. On the other hand, expert features are developed to describe particular movement patterns highlighted in earlier studies. Among these, two expert features — static periods and movements falling below a certain threshold — were computed. Both, generic and expert features, are listed in Table 3.5.

Below Threshold

This metric counts the number of samples falling under a specific threshold, which, following the methodology described in previous research [Ric22], is determined as a percentage of the signal's maximum value. This metric serves as an additional method to assess expected stationary behavior under acute stress conditions, with the threshold ratio defined at 10%, as detailed in Equation 3.5.

Static Periods

As in previous work, static periods were calculate to assess if participants are moving less under the influence of acute stress [Ric22]. Static periods were identified across 0.5 second windows with a 50% overlap. A window was classified as static if its variance fell below 10% of signal's maximum value. This threshold was chosen by using test data.

Table 3.5: Overview of both generic and expert movement features for each body part.

Body Part	Type	Metric	Definition
left/ right shoulder	generic	mean_velocity	Mean velocity
		std_velocity	SD velocity
		range_of_motion	Range of motion
head	expert	below_threshold	Counts per minute below threshold
		static_periods	Counts per minute of static periods
left/right hand	generic	visibility	Absolute count of visibly detected hands
left/right elbow	generic	visibility	Absolute count of visibly detected elbows

$$\text{below\_threshold}(x) = \begin{cases} \text{True}, & \text{if } \|x\|_2 \leq 0.1 \cdot \max(x) \\ \text{False}, & \text{otherwise} \end{cases} \tag{3.5}$$

Gaze Behavior

The evaluation of gaze behavior utilizes eye-specific landmarks, which are selectively extracted from the facial landmarks previously used in head movement analysis with the $\mathrm{MediaPipe}$ package. Isolating gaze movements from head motions requires subtracting the mean absolute head movement from these eye landmarks. This process enables the calculation of mean velocity, standard deviation, and the identification of periods below threshold and static, mirroring the analysis of movement features described earlier. A comprehensive list of gaze-centric features is provided in Table 3.6.

Focusing on changes in pupil diameter due to acute stress, only landmarks related to the pupil are extracted. The $\mathrm{smallestenclosingcircle}$ package, based on the work by Welzl [Wel91], is used to draw a circle around these landmarks for both the left and right eyes, which facilitates the calculation of the diameter. Based on these diameters, general statistics such as mean, standard deviation, and the minimum and maximum values are calculated, as depicted in Table 3.6.

To study blinking patterns, distances between the eyes' vertical and horizontal landmarks are calculated frame by frame. A blink is identified when the Eye Aspect Ratio (EAR) is below a predetermined threshold of 0.47, based on test data analysis. The blink rate for each eye is then quantified, adjusting the count to a per-minute rate.

Table 3.6: Overview of features for gaze behavior, pupil diameter and blinking behavior.

Eye Side	Type	Metric	Definition
left	gaze	mean_velocity	Mean gaze velocity
		std_velocity	SD gaze velocity
		below_threshold	Counts per min below threshold
		static_periods	Counts per min static periods
right	pupil	mean_pupil	mean pupil diameter
		std_pupil	SD pupil diameter
		min_pupil	minimal pupil diameter
		max_pupil	maximal pupil diameter
		diff_pupil	framewise difference in pupil diameter
	blinks	blinks_min	mean blinks per minute

3.2.3 Aims

Aim 1: To predict acute stress versus a non-stressful control condition using video-based digital biomarkers and to analyze how facial expressions, body movements, and gaze behavior contribute to the classification output.

- **Type**: Inference-based Statistics, ML-based Classification
- **Hypothesis**: Video-based digital biomarkers predict stressed (TSST) and non-stressed (f-TSST) states with high discriminatory accuracy. Facial expressions, body movements, and gaze behavior each influence the classification output.

Aim 2: To examine the relationship between changes in facial expressions, speech, and movement, and traditional biological and psychological stress markers.

- **Type**: ML-based Regression
- **Hypothesis**: Digital biomarker predict continuous values of HR, cortisol, and subjective stress scores.

3.2.4 Model Development & Evaluation

To assess how effective digital biomarkers from video recordings are at identifying acute stress, several different models were employed.

Inference-based Statistics

Given the sample size exceeds 30, the distribution of the samples is assumed to be normal in accordance with the central limit theorem [Kwa17]. Consequently, the statistical analysis employed a repeated measures Analysis of Variance (ANOVA) for the main test, complemented by paired t-tests as post-hoc analysis. Since participants were subjected to both control and intervention conditions, a within-subjects design was chosen. Statistical computations were conducted using the Python package biopsykit [Ric21b], leveraging pingouin [Val18] for implementation. Effect sizes for t-tests were quantified using Hedge's g. Statistical significance was denoted in Figures and Tables as follows: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

ML-based Classification

ML models were developed to distinguish between stress and non-stress states using digital biomarkers. The classification process followed a standard pipeline, comprising of removing features with zero variance, scaling, feature selection, and classification, as further detailed in Figure 3.4. To determine the best pipeline with respect to classification performance, different combinations were tested which will be described in the following.

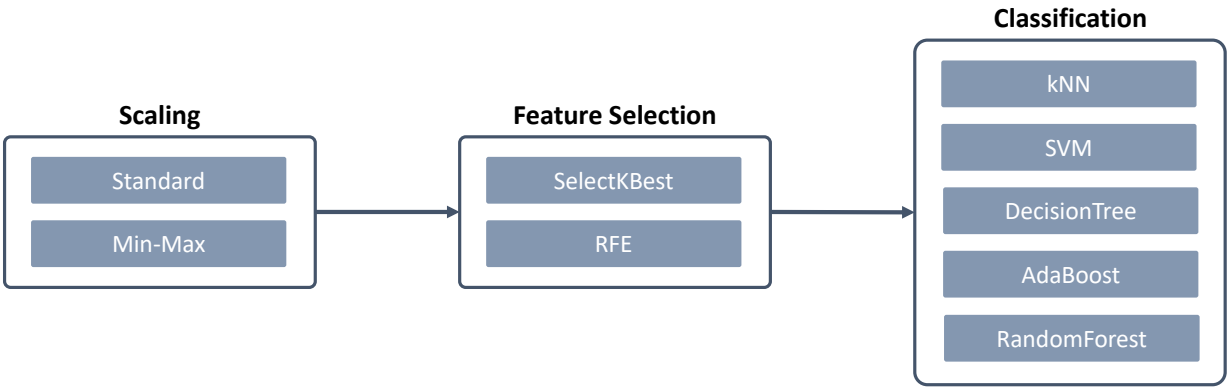

Figure 3.4: Standard ML-based classification pipeline.

First, all features with zero variance were removed. For scaling, both the Min-Max Scaler and the Standard Scaler were tested. The Standard Scaler adjusts features to have a mean of 0 and a standard deviation of 1 through z-score normalization, while Min-Max Scaler normalizes features to a range of 0 to 1. Following scaling, feature selection was performed using either Select-k-Best (SkB) or Recursive Feature Elimination (RFE). SkB selects the top k features based on the highest ANOVA F-scores, whereas RFE uses an estimator (in this case, Support Vector Machine (SVM)) to rank features by importance and recursively eliminate the least important

until the specified feature count is met. For classification, five distinct models were compared based on their maximal accuracy: k-Nearest-Neighbors (kNN), Decision Tree (DT), SVM, and RandomForest classifiers.

All different pipeline combinations were tested using five-fold cross-validation (CV). Within each CV fold, another five-fold CV was used for feature scaling and hyperparameter optimization to determine the best combination of feature selection algorithms and classifiers. Therefore, the inner CV will be referred to as *hyperparameter search CV* and the outer one as *model evaluation CV*. To test all possible hyperparameter combinations, the accuracy was used as a target metric and a grid search was employed for both the feature selection process and the classifiers, with random search employing 10,000 trials specifically for Random Forest (RF).

Afterwards, the pipelines were retrained with the hyperparameters that yielded the highest accuracy on all input data of the hyperparameter search CV to evaluate the different classification pipelines. Predictions were then performed on the test set of the model evaluation CV, which the classifier had not seen yet. This process was repeated for each fold of the model evaluation CV. To avoid possible train-test leaks, it was ensured that data from the same participant (TSST and f-TSST) were either present in the training or the test set for both the hyperparameter search CV and the model evaluation CV.

The performance for each pipeline combination was evaluated over all folds of the model evaluation CV. Therefore, the classification metrics accuracy, F1-score, and precision were computed and confusion matrices were derived. Further, SHapley Additive exPlanation (SHAP) values, an explainable machine learning approach, were employed using the `shap` Python Package [Lun17], to identify the key features responsible for influencing individual outcome predictions [Lun20].

The complete list of hyperparameters used for optimization is detailed in Table 3.2.4.

Backward linear regression

To relate digital biomarkers to HR, cortisol levels, and self-reported scores, stepwise backward multiple regression (SBMLR) was explored, following the methodology of Lasselin et al. [Las20]. Consequently, the input data needs preparation: First, the difference between f-TSST and TSST was calculated, followed by the performance of z-score normalization for comparability. First, features exhibiting multicollinearity greater than 0.8 were excluded to ensure model reliability. Secondly, to reduce complexity and the dimensionality of the digital biomarkers, PCA was applied individually to the three feature groups facial expressions, movement, and gaze patterns, retaining only the components that accounted for at least 80% of the variance.

Regarding SBMLR model development, the goal was to refine the model to achieve the optimal balance between explanatory power, as indicated by the highest adjusted R^2 value [Chi21], and model simplicity. Through an iterative process, the SBMLR method selectively removed the least significant predictor at each step. Based on the highest adjusted R^2 value, the final best-fitting model was selected.

ML-based Regression

Regression analysis, employing ML techniques, was conducted similarly to the classification approach. Unlike the classification models, which identify condition labels, the regression models were designed to predict continuous endocrine (m_{S1S4}) and self-reported measures (PANAS Negative and Positive Affect), as well as HR levels (mean HR). Consequently, five specific models were adapted to their regression counterparts: kNN Regressor, Support Vector Regressor (SVR), Decision Tree Regressor, AdaBoost Regressor, and RandomForest Regressor. The models were optimized based on the metric R^2 (coefficient of determination). Hyperparameter optimization was performed using a nested five-fold cross-validation technique in combination with a grid search.

The regression pipeline employed the same feature scaling and selection techniques as the classification task, including the application of either StandardScaler or MinMaxScaler for scaling, and SelectKBest or RFE for feature dimensionality reduction.

The hyperparameter configurations for these regression models are detailed in Table 3.2.4.

The evaluation of the best ML-based regression model relied on two metrics: the R^2 value, indicating the model's explanatory power, and the MAE, which reduces outlier influence, ensuring a robust assessment [Chi21].

Table 3.7: Hyperparameter grid used for classification and regression.
[1] only for RBF kernel; [2] only for poly kernel; [3] RandomizedSearch was used for Randomforest

Feature Selection	**Hyperparameter**	**Values**
SelectKBest	k	2 to 30; steps of 2
RFE	n	2 to 20; steps of 2
Classifier/Regressor	**Hyperparameter**	**Values**
kNN	k	1 to 20; steps of 2
	weights	uniform, distance
SVM	Kernel	linear, RBF, poly
	C	10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}
	gamma[1]	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}
	degree[2]	2 to 6
DecisionTree	criterion classification	gini, entropy
	criterion regression	squared_error, friedman_mse
	depth	2 to 20; steps of 2
	min_samples_leaf	0.1 to 0.5; steps of 0.1
	min_samples_split	0.1 to 0.8; steps of 0.1
	max_features	0.1 to 0.6; steps of 0.1; $sqrt$, log_2, all
AdaBoost	base_estimator	DecisionTree
	n_estimators	10 to 400; steps of 30
	learning_rate	0.001, 0.01, 0.1, 1.0
RandomForest [3]	bootstrap	True, False
	criterion classification	entropy
	criterion regression	squared_error, friedman_mse
	max_depth	4 to 50; steps of 2; all
	max_features	0.1 to 0.5; steps 0.1; $sqrt$
	min_samples_leaf	0.05 to 0.2; steps of 0.05
	min_samples_split	0.1 to 0.6; steps of 0.1
	min_weight_fraction_leaf	0.0 to 0.5; steps of 0.1
	max_leaf_nodes	2 to 20; steps of 2
	min_impurity_decrease	0 to 0.1; steps of 0.01
	n_estimators classification	10 to 500; steps of 20
	n_estimators regression	10 to 400; steps of 40
	ccp_alpha	0 to 1; steps of 0.1

3.3 rPPG

To validate state-of-the-art rPPG models, this chapter outlines the utilized models, including both conventional and DL-based models, as well as benchmark datasets. Further, the rPPG pipeline will be described, including its pre-processing steps, model integration, and HR extraction from the predicted rPPG signal. An overview of the validation process, encompassing both training and testing datasets, as well as evaluation metrics, will then be provided. As this thesis aims to assess the performance of current rPPG models, only openly available pre-trained models have been utilized.

3.3.1 Conventional Models

This section delves into the conventional rPPG methods that have set the stage for remote HR monitoring techniques from video recordings. A summary of the methods used in this thesis is provided in Table 3.8.

Table 3.8: Overview of conventional rPPG methods utilized in this thesis.

Method	Year	Description
GREEN [Ver08]	2008	Initial demonstration of the feasibility of the rPPG method, revealing that the green channel contains the strongest pulsatile signal.
ICA [Poh10]	2010	ICA-based rPPG isolates the BVP signal from RGB video data by decomposing it into independent components, removing correlations and noise sources.
CHROM [De 13]	2013	CHROM mitigates motion artifacts by employing linear combinations of RGB channels, leveraging skin color changes from diffuse and specular reflections to isolate pulse signals.
POS [Wan17]	2017	POS isolates HR signals in rPPG by using the orthogonal plane to skin tone in RGB space, merging weighted, normalized channels to separate physiological signals from noise.
LGI [Pil18]	2018	LGI stabilizes HR estimation in rPPG by using local group invariance to filter out motion and lighting disturbances, focusing on invariant physiological signals.

GREEN

The GREEN method was proposed by Verkruysse et al. [Ver08] and is the first approach in rPPG. To extract the HR, it leverages the green channel of RGB video data, exploiting hemoglobin's higher sensitivity to green light for stronger pulsatile signal detection. This technique involves capturing the green channel intensity fluctuations over time, filtering to isolate the HR signal from noise, and calculating the HR by identifying the peak frequency in the signal's Fourier transform.

ICA

ICA is a statistical technique that separates a multivariate signal into additive, independent non-Gaussian components. In the context of rPPG, this ICA method [Poh10] decomposes RGB color signals from video into independent sources, employing an algorithm that utilizes the joint approximate diagonalization of eigenmatrices to remove correlations and high-order dependencies within the RGB channels. The premise is that the BVP signal can be isolated as one of these independent components, distinct from other noise sources such as ambient light variations and motion artifacts.

CHROM

CHROM [De 13] is a rPPG method which exploits the BVP feature to discriminate pulse signals from motion distortions by using both diffuse and specular reflections. These reflections cause the observed skin color to change based on the angle and distance between the camera, the skin, and the light sources. By adopting the CHROM technique, it becomes possible to mitigate the impact of motion artifacts through the use of linear combinations of the red, green, and blue color channels under a standardized skin-tone assumption.

POS

The POS method [Wan17] in the context of rPPG is a sophisticated technique designed to extract HR signals by utilizing the orthogonal plane to the skin tone in RGB signal space. By combining and normalizing RGB channels into two distinct channels, and then weighting these to merge into a singular, optimized signal, the POS method effectively isolates the physiological rPPG signals from ambient noise and variations in skin tone.

LGI

The Local Group Invariance (LGI) method by Pilz et al. [Pil18] improves HR estimation from face videos under variable conditions by using mathematical invariance principles. LGI extracts robust features that are less affected by common disturbances like motion and lighting changes. This method reorganizes the blood volume signal in a vector space, concentrating its distribution for better accuracy. Proven effective in challenging scenarios, LGI outperforms traditional rPPG techniques by maintaining high accuracy across different environmental and motion conditions.

3.3.2 Deep Learning Models

This section discusses five distinct DL rPPG methods, with a focus on validating pre-trained models. Therefore, only models that are pre-trained and openly available were considered. An overview is provided in Table 3.9.

Table 3.9: Overview of DL rPPG methods utilized in this thesis.

Model	Year	Modules	Description
DeepPhys [Che18b]	2018	2D CNN	Enhances rPPG extraction with attention-enhanced motion and appearance models, offering improved robustness but limited temporal analysis.
TS-CAN [Liu21a]	2021	2D CNN, TSM	Improves rPPG extraction with a Temporal Shift Module, enabling enhanced temporal analysis compared to DeepPhys.
Efficient-Phys [Liu21b]	2021	2D CNN, TSM, Attention Module	Streamlines physiological measurement with a CNN approach, eliminating pre-processing and focusing on simplified, direct video analysis.
PhysNet [Yu19]	2019	3D CNN	Learns temporal and spatial contextual features using a 3D CNN backbone architecture to predict rPPG based on raw RGP input videos.
Phys-Former [Yu22]	2022	Transformer, 2D CNN	Employs a time-difference transformer approach for advanced rPPG signal analysis, emphasizing enhanced spatiotemporal feature refinement.

DeepPhys

DeepPhys [Che18b], an innovative model inspired by HR-CNN [Spe18] (first DL-based rPPG model), advances the extraction of rPPG signals through a 2D CNN. This novel approach differentiates itself by incorporating both a motion model and an appearance model, based on the DRM. The key innovation lies in the appearance model's use of an attention mechanism to guide the motion model in learning representations, utilizing the normalized differences between adjacent frames to simulate motion and color changes. This enhances the model's robustness to motion. Furthermore, DeepPhys introduces an attention mechanism for generating soft attention masks from video frames, assigning greater importance to skin regions exhibiting stronger physiological signals and visualizing the spatiotemporal patterns of these signals. While DeepPhys shows improved performance over HR-CNN, especially in conditions of lighting changes and motion artifacts, it is limited by the inability of 2D CNNs to fully capture the temporal dynamics of rPPG signals. The architecture of DeepPhys is shown in Figure 3.5.

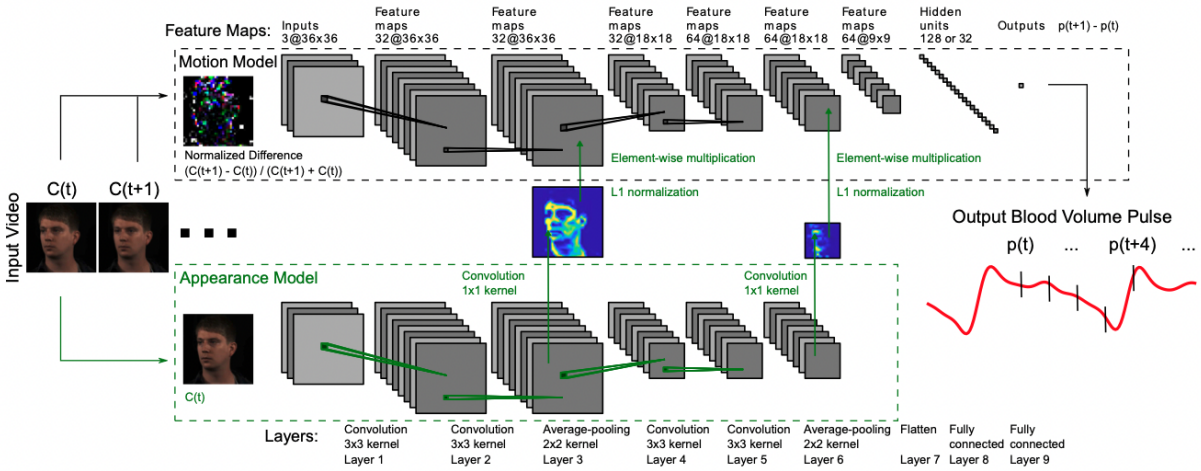

Figure 3.5: DeepPhys: Architecture of the end-to-end CNN, processing current and differential video frames to learn shared spatial masks and features for BVP and respiration signal recovery. [Che18b].

TS-CAN

TS-CAN [Liu21a] builds on the foundation laid by DeepPhys, enhancing the capture of temporal information in rPPG signal extraction through the integration of a Temporal Shift Module (TSM). This addition facilitates the exchange of information between adjacent frames without resorting to complex convolution operations, by shifting blocks in tensors along the time axis. This mechanism

allows TS-CAN to effectively incorporate temporal dynamics into its analysis. Diverging from DeepPhys, TS-CAN's appearance model inputs are frames averaged from multiple adjacent frames rather than raw video frames, a strategy that ensures the model captures temporal information more effectively. The model is shown in Figure 3.6.

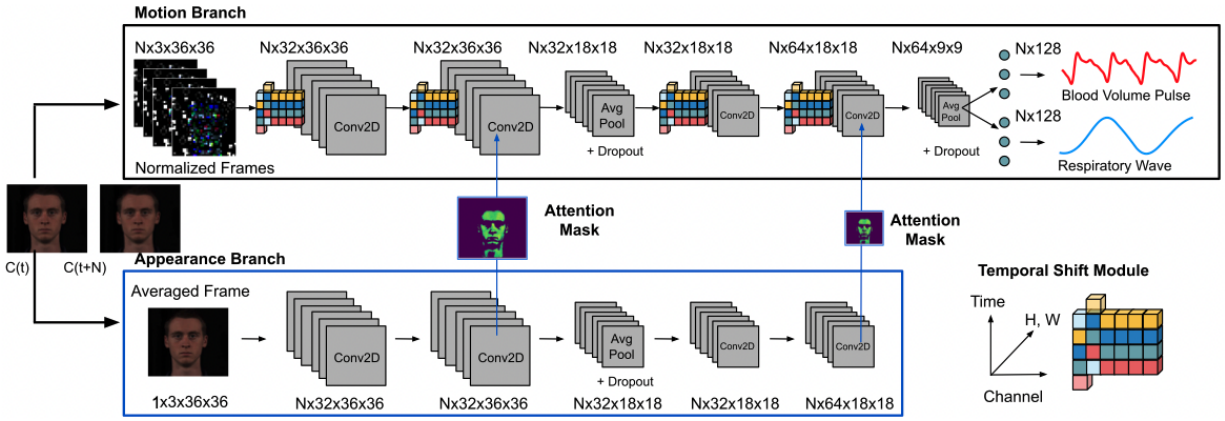

Figure 3.6: TS-CAN: End-to-end temporal shift convolutional attention network for camera-based physiological measurement [Liu21a].

EfficientPhys

EfficientPhys [Liu21b] introduces a CNN-based method for camera-based physiological measurements, closely resembling the TS-CAN model. However, EfficientPhys distinguishes itself by eliminating the necessity for pre-processing steps such as face detection or color space transformation. The novel method directly processes raw video frames through a streamlined single-branch architecture equipped with a custom normalization layer and a self-attention mechanism, visualized in Figure 3.7. Similar to TS-CAN, it employs a TSM, and a 2D convolution process. This setup enables effective and accurate spatial-temporal analysis while ensuring ease of deployment due to its simplicity.

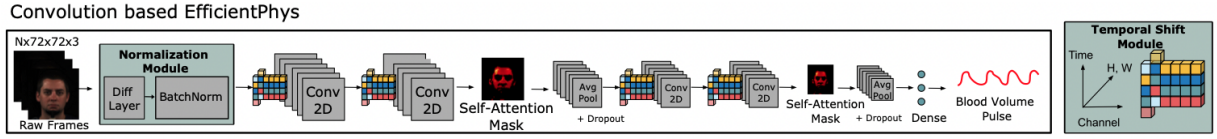

Figure 3.7: EfficientPhys: Single-branch CNN with custom normalization layer, self-attention mechanism, and TSM [Liu21b].

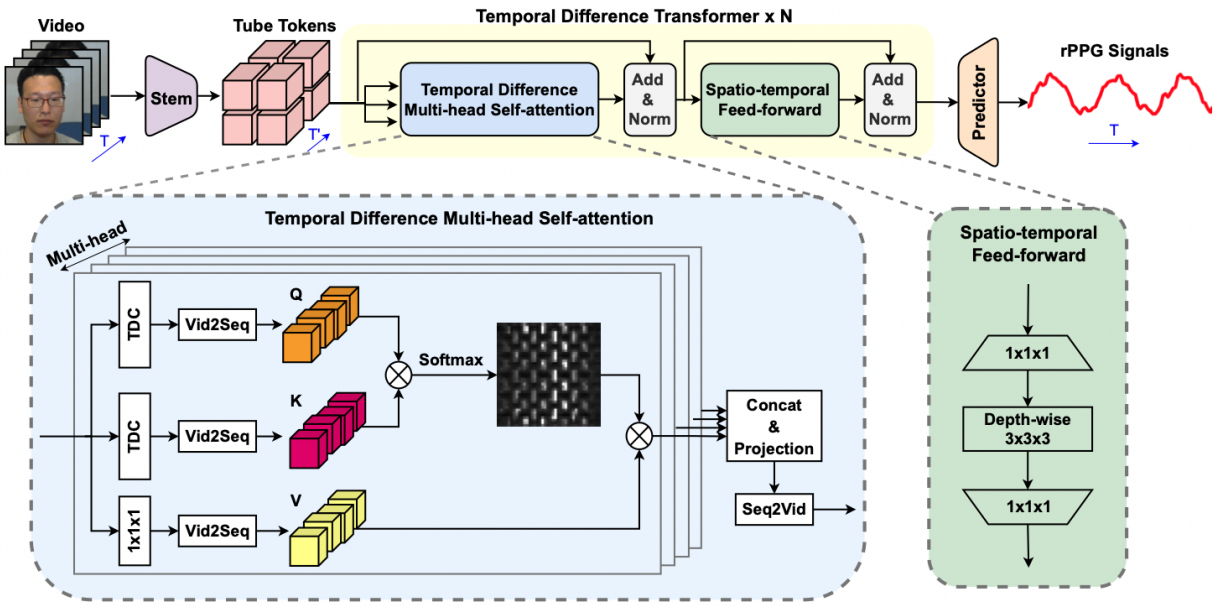

Figure 3.8: PhysFormer: Features a shallow stem, tube tokenizer, temporal difference transformers with TD-MHSA and ST-FF modules for improved spatio-temporal analysis, and an rPPG predictor. TDC refers to temporal difference convolution [Yu22].

PhysNet

PhysNet, proposed by Yu et al. [Yu19], builds on a 3D CNN backbone network and does not rely on any pre-processing operations. It directly inputs the raw RGB video frames into the 3D CNN network. PhysNet simultaneously learns temporal and spatial features of facial sequences to achieve more robust context recognition and reduce temporal fluctuation in recovering rPPG signals. In their method, videos are treated as consistent collections of frames, and the raw video streams are directly input into the 3D CNN backbone network without any image pre-processing steps, such as face detection.

PhysFormer

Yu et al. introduced PhysFormer [Yu22], a novel approach leveraging the transformer architecture for HR measurement, marking a significant advancement in rPPG techniques. PhysFormer is distinguished by its utilization of a time-difference transformer to capture long-range spatiotemporal relationships within the rPPG signal estimation process. This method excels in encoding local time differences and facilitating global spatiotemporal analysis, effectively guiding the attention mechanism towards quasi-periodic rPPG features while enhancing local spatiotemporal representations, especially in scenarios with interference. Initial Region of Interest (ROI) selection on RGB facial

video inputs is conducted using a Multi-task Cascaded Convolutional Networks (MTCCN) based face detector. The process generates spatiotemporal tags, serving as input for the time-difference transformer. Rather than directly producing rPPG signals, PhysFormer processes these features through an rPPG prediction head for temporal upsampling and spatial averaging. This step further refines the features, culminating in the generation of 1D rPPG signals. The model's architecture is shown in Figure 3.8.

3.3.3 Datasets

This section examines the benchmark datasets that are critical for rPPG research. A recent comprehensive review in the field identified the UBFC-rPPG, PURE, and COHFACE datasets as the most frequently utilized in rPPG studies [Xia24]. Given one aim of the thesis is to validate commonly used pre-trained rPPG models, these datasets have been included into this thesis. Table 3.10 provides an overview of the datasets utilized in this thesis. Additionally, the HR ranges of the different datasets are shown in Figure 3.9.

Table 3.10: rPPG datasets used in this thesis (#P: Number of Participants, #V: Number of Videos)

Dataset	#P	#V	Measure	Illumination	Movement	HR range
UBFC-rPPG	42	42	HR, PPG	studio lighting natural lighting (windows)	static	56-130 bpm
PURE	10	60	HR, SpO2, PPG	natural light shadow effects	6 diff. head movements	46-132 bpm
COHFACE	40	160	BVP, RR	studio lighting natural lighting (windows)	static	47-70 bpm
UBFC-Phys	56	168	PPG, HR	studio lighting	naturalistic	45-135 bpm
EmpkinS-TSST	44	264	ECG	studio lighting	naturalistic	46-161 bpm

PURE

The PURE dataset [Str14] is composed of video recordings from 10 individuals, encompassing 8 men and 2 women. These participants each produced 6 videos, resulting in a collective total of 60 videos. The videos were captured at a resolution of 640x480 pixels and a frame rate of 30 frames per second (fps), each lasting for one minute. To induce variability in head movement, participants

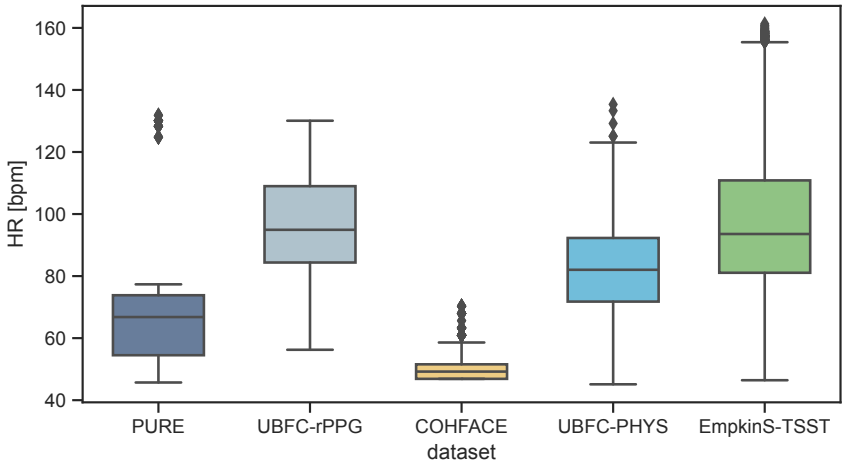

Figure 3.9: HR ranges of the video datasets PURE, UBFC-rPPG, COHFACE, UBFC-PHYS and EmpkinS-TSST

were instructed to engage in six distinct activities: (1) remaining stationary, (2) speaking, (3) executing slow head rotations, (4) conducting rapid head movements, (5) turning their head to a 20-degree angle, and (6) angling their head at 35 degrees. Additionally, the PURE dataset accounted for variations in lighting conditions by utilizing natural light and shadow effects through a substantial window. Authentic PPG signals were obtained using a CMS50E finger pulse oximeter at a 60 Hz sampling frequency. Notably, the PURE dataset's images are preserved in the lossless PNG format, enhancing the accuracy of rPPG signal estimation.

UBFC-rPPG

The UBFC-rPPG dataset [Bob19] has been specifically designed to benchmark rPPG methodologies. It encompasses 50 videos, each featuring a unique individual, captured at a resolution of 640x480 pixels and a frame rate of 30 fps, with an emphasis on capturing varying lighting conditions, from sunlight to indoor lighting. The UBFC-rPPG dataset is divided into two segments. The first segment comprises 8 videos in a controlled setting where participants were instructed to remain motionless, however incidental movements occur in some recordings. The second segment, aimed at mimicking real-world conditions, contains 42 videos of participants engaged in a time-sensitive mathematical challenge designed to elevate their HR. The UBFC-rPPG stands out as a frequently utilized resource in research due to its uncompressed, high-quality video footage and the inclusion of authentic HR and PPG data. While it consists of two segments, the practicality and superior video quality of the second segment was used in this thesis.

UBFC-Phys

The UBFC-Phys dataset [Sab23] is specifically designed for the study of emotion recognition, featuring 56 participants with 46 women and 10 men. Using the TSST, participants performed three tasks: resting, speaking, and performing arithmetic tasks. This setup generated a total of 168 videos, with each video captured at a high resolution of 1024x1024 pixels and a frame rate of 35 fps. For physiological data collection, the Empatica E4 wristband was employed to gather PPG signals and EDA. Furthermore, to assess self-reported levels of anxiety, participants completed questionnaires which are further described in the paper by Sabour et al. [Sab23].

COHFACE

The COHFACE dataset [Heu17], developed by the Idiap Research Institute, is a resource made publicly available for the purpose of allowing researchers to test their rPPG methodologies under uniform and equitable standards. The dataset encompasses recordings from 40 subjects, with a distribution of 28 males and 12 females. Each participant contributed to the creation of four video clips, resulting in a total of 160 videos in the dataset. These videos were captured at a resolution of 640x480 pixels and a frame rate of 20 fps. Participants were equipped with a contact-based PPG sensor during the recordings. The dataset also carefully considers lighting conditions by recording two sets of video clips for each participant under distinct lighting scenarios: (1) controlled studio lighting achieved with window blinds to block natural light while providing adequate artificial illumination for consistent facial exposure; and (2) natural lighting conditions, where recordings were made with any window blinds and all artificial lighting sources switched off. A notable challenge with COHFACE is its extensive video compression, which has implications for the quality of the data provided.

3.3.4 rPPG Pipeline

In this thesis the rPPG toolbox was used to validate conventional (Section 3.3.1) and DL-based rPPG models (Section 3.3.2). The toolbox employs a configuration file system, enabling users to modify parameters for pre-processing, training, post-processing, and evaluation via a YAML file for each experiment. This system offers detailed control over hyperparameters and computational resources, with both neural and unsupervised methods sharing settings like input resolution and face cropping. In Table 3.11, the hyperparameter configuration of the pre-trained models is depicted. In the following each step of the pipeline will be outlined.

Table 3.11: rPPG datasets used in this thesis (#P: Number of Participants, #V: Number of Videos)

Model	Input	Chunk	Data Input	Label Input	Crop Face
GREEN, ICA, CHROM,BVP, POS	72x72	210	[Raw]	Raw	Yes
DeepPhys	72x72	180	[DiffNormalized, Standardized]	DiffNormalized	Yes
TS-CAN	72x72	180	[DiffNormalized, Standardized]	DiffNormalized	Yes
EfficientPhys	72x72	180	[Standardized]	DiffNormalized	No
PhysNet	72x72	128	[Raw]	DiffNormalized	No
PhysFormer	128x128	160	[DiffNormalized]	DiffNormalized	Yes

Pre-processing

The pre-processing pipeline involves several steps to bring the video input and ground truth PPG label of the different datasets, which were detailed in Section 3.3.3, in the desired shape. Initially, the PPG signal is resampled to ensure it matches the length of the video. Following this, video inputs undergo face cropping to separate the face from the background using OpenCV's Haar Cascade classifier [Vio01]. To avoid excluding important facial features, a bounding box with a scale factor of 1.5 is employed, with the cropping process applied individually to each frame. Given that each model requires specific input sizes, the cropped faces are resized accordingly. The images of the videos and labels are then processed based on the data type: they are either differentially normalized — calculating the discrete difference along the time axis in video data/label and normalizing by its standard deviation — or z-score standardized. Due to the models' inability to process entire video sequences simultaneously, the videos are segmented into chunks, with lengths specific to each model. It is important to note that for EfficientPhys and PhysNet, both end-to-end models, face cropping is not conducted.

Model Training & Testing

Following the pre-processing phase, the processed data was fed into one of the rPPG models, as detailed in Sections 3.3.1 and 3.3.2, to predict the PPG signal from the input video. During the training process, the developers of the rPPG-toolbox adopted a consistent approach, using a batch size of 4 and conducting training over 30 epochs, while also setting an inference batch size of 4

for all models. Additionally, they customized various loss functions, optimizers, and learning rate schedulers specifically for each DL-based model to optimize performance. A detailed summary of the hyperparameter settings for the pre-trained models is available in the Appendix.

Post-processing

In the post-processing phase, the rPPG-toolbox implements standard procedures to refine each model's PPG prediction, which includes filtering and smoothing the signal, followed by HR estimation using fast Fourier transform (FFT). These steps are applied to the predicted signal chunks, each covering a duration of 5-7 seconds. The choice of window size in FFT analysis has a direct impact on frequency resolution and the ability to distinguish between close frequencies, affecting the accuracy of heart rate calculation. The frequency resolution in bpm Δf_{bpm} can be calculated with the following Formula:

$$\Delta f_{bpm} = \frac{60}{T_s} \tag{3.6}$$

Following Formula 3.6, a window size of 5-7 seconds would result in a frequency resolution of 12-8.5 bpm. Increasing the window size improves frequency resolution but decreases time resolution, making it harder to track rapid changes in heart rate over time. Therefore, in this thesis a window size of 20 seconds was chosen, which results in a frequency resolution of 3 bpm. To address the time resolution, this thesis uses a window-based approach for HR estimation across longer signal segments. This method involves first aggregating the predicted PPG chunks from each video into a unified signal. This combined signal is then refined with a 2nd-order Butterworth filter, employing cut-off frequencies between 0.75 and 2.5 Hz, to smooth the signal. Following this, HR is determined on a secondly basis within a 20-second window, utilizing either the peak detection algorithm or FFT. These post-processing steps applied to the predicted PPG signal result in an HR estimation for each second of the input video.

3.3.5 Aims

Aim 3: To determine whether the performance of conventional and DL rPPG methods decreased on more naturalistic datasets which address factors such as varying heart rate levels, speech sequences, and head movements.

- **Type:** Statistical Analysis
- **Hypothesis:** Both conventional and DL-based rPPG models will perform worse on more naturalistic datasets that include speaking, head movements, and higher heart rate levels, compared to their performance on controlled laboratory benchmark datasets.

3.3.6 Model Development & Validation

Each conventional and DL-based model were tested against each other to determine the best performning model on the various benchmark and EmkinS-TSST dataset. The DL-based models were pre-trained on the following datasets: UBFC-rPPG and PURE. For evaluation, the models were first cross-tested on PURE and UBFC-rPPG. Then the models were validated on the datasets COHFACE, UBFC-Phys and EmpkinS-TSST were used.

3.3.7 Evaluation

Statistics

To compare the predicted HR levels with the ground truth across different rPPG models, the following measures were utilized: Mean absolute error (MAE), root mean squared error (RMSE), Mean absolute percentage error (MAPE), and Pearson Correlation (ρ).

Bland-Altman Plot

To assess the accuracy of various rPPG models, the Bland-Altman plot is utilized to compare differences between paired measurements from rPPG algorithms to their average values. This method evaluates systematic bias and agreement between algorithms, showing bias through the average difference and reliability through limits of agreement—calculated as ± 1.96 times the standard deviation around the average difference. A narrower range indicates higher concordance among algorithms, demonstrating their effectiveness for accurate, non-invasive vital sign monitoring [Bla86]

3.4 Multimodal Stress State Detection

In its final objective, this thesis assesses the effectiveness of multimodal stress detection through the integration of digital biomarkers and rPPG. It expands upon stress state detection by incorporating HR predictions derived from rPPG, now referred to as rPPG-derived HR (rPPG-HR), aiming to differentiate between stressed and non-stressed states within the EmpkinS-TSST dataset. The subsequent sections will detail the input features, expected outcomes, and the development of the model employed in this analysis. An overview is given in Figure 3.10.

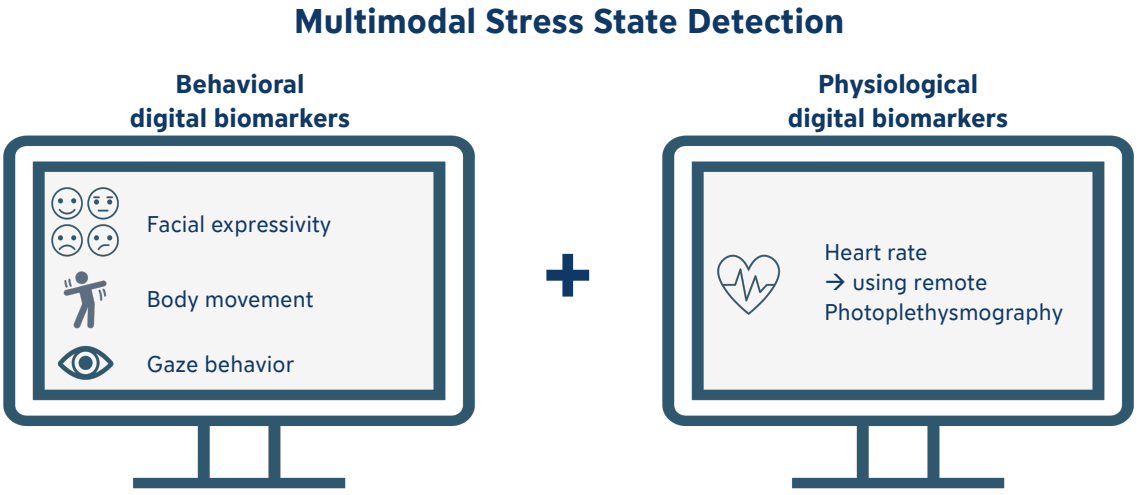

Figure 3.10: Multimodal stress state detection: Extending the digital biomarkers by incorporating HR predictions derived from rPPG, aiming to differentiate between stressed and non-stressed states within the EmpkinS-TSST dataset.

3.4.1 Features

For multimodal stress state detection on the EmpkinS-TSST dataset, digital biomarkers, meaning facial expressions, body movements, and gaze behavior, were combined with rPPG-derived HR (rPPG-HR). A detailed extraction of the digital biomarkers can be found in Section 3.2.2. For the rPPG-HR, the outcome of the best performing rPPG on the EmpkinS-TSST was utilized, which was described in 3.3.

3.4.2 Aims

Aim 4: To examine whether the combination of behavioral digital biomarkers with rPPG-derived HR (rPPG-HR) increases the stress prediction accuracy compared to behavioral digital biomarkers alone.

- **Type:** ML-based Classification
- **Hypothesis:** The combination of behavioral digital biomarkers with rPPG-derived HR (rPPG-HR) improves the prognostic performance for stress prediction as compared to behavioral digital biomarkers alone.

3.4.3 Model Development & Evaluation

Inferential Statistics

Inferential statistics were used to evaluate condition-related differences in the rPPG-HR, following the methods outlined in Section 3.2.4.

ML-based Classification

As previously described in Section 3.2.4, a ML classification pipeline has been established to classify stress states exclusively through facial expressions, body movements, and gaze behavior. In this study, the same pipeline is leveraged to address a different question: Identifying the best-performing classifier for multimodal stress state detection.

Chapter 4

Results

4.1 Stress Assessment

The following Sections show the results of traditional stress markers like salivary cortisol, HR(V) measures, and questionnaires such as PANAS and State-Trait Anxiety-Depression Inventory (STADI).

4.1.1 Saliva

Overall, cortisol levels increased by 159.73% from S_1 to the anticipated highest peak at S_3, as depicted in Figure 4.1. Conversely, in the f-TSST, the increase was only 37.51%. A significant difference was observed in all derived features when comparing the TSST to the f-TSST. Notably, the largest effect sizes were observed for m_{S1S4} with a Hedges' g value of 0.808 and Δc_{max} with g = 0.745, as detailed in Table 4.1. These features are further illustrated in Figure 4.2.

Table 4.1: t-test results of cortisol features; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Feature	t(43)	p	Hedges' g
AUC_G	3.097	0.019*	0.556
AUC_I	3.766	0.003**	0.710
Δc_{max}	3.673	0.004**	0.745
m_{S1S4}	3.883	0.002**	0.808

Regarding the condition order, a more pronounced effect was noted when the TSST was administered first, with m_{S1S4} showing a Hedges' g value of 1.160 and Δc_{max} at 1.010. In contrast, when the f-TSST was the initial condition, m_{S1S4} and Δc_{max} showed smaller effect sizes with

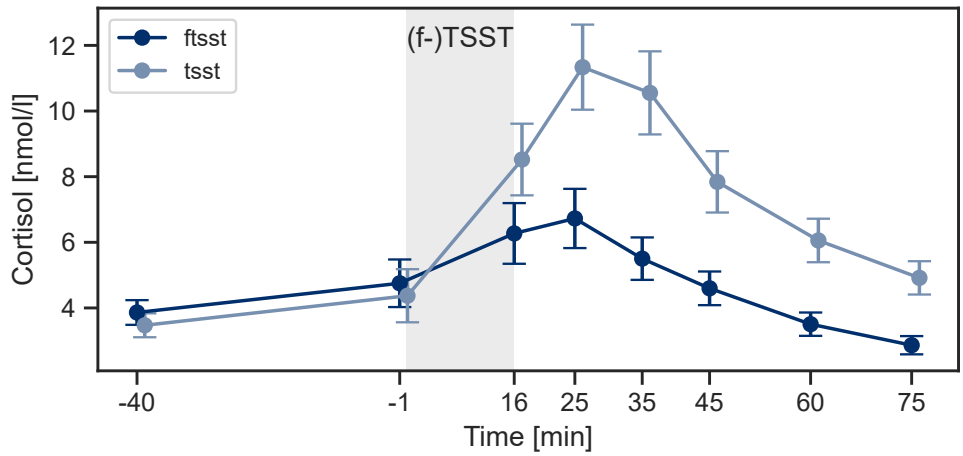

Figure 4.1: Cortisol response: Mean±SE over all participants

Hedges' g values of 0.485 and 0.486. The cortisol responses during the (f-)TSST are also visualized in greater detail in the Appendix A, particularly in Figure A.1 for the condition order, in Figure A.2 for gender differences and in Figure A.3 for the standing and sitting condition.

For sAA, no significant differences were found between TSST and f-TSST condition. The results are shown in Appendix A in Figure A.4.

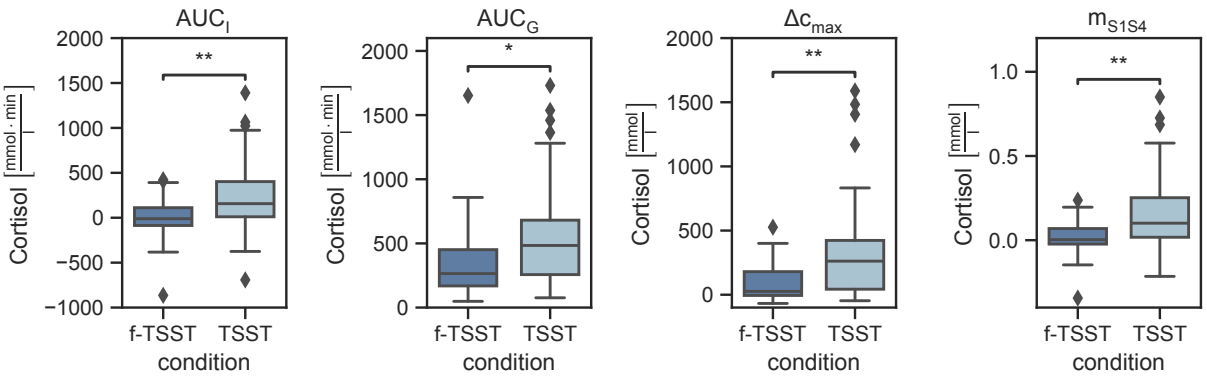

Figure 4.2: Features derived from cortisol over all participants; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

4.1.2 Heart Rate

The overall mean HR was significantly elevated during the TSST compared to the f-TSST, as shown in Figure 4.3. In terms of HRV metrics, meanNN, SDNN, RMSSD and pNN50 were significantly higher in the TSST. This pattern of significance was consistent across the different

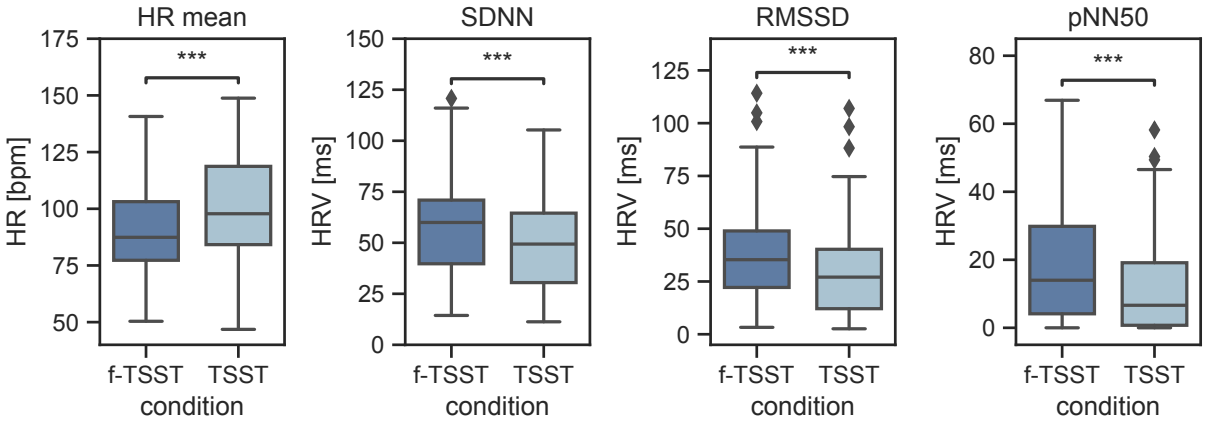

Figure 4.3: HR and HRV results over all participants; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

phases of the experiment for both mean HR and HRV measures, as further detailed in the Appendix in Figure A.5.

Upon comparing the sitting and standing conditions, the standing condition exhibited a greater effect size with a Hedges' g = 0.570 for mean HR as opposed to g = 0.418 for the sitting condition. Detailed statistical analyses, including all t-tests, are presented in Table B.6 in the Appendix B. Regarding the order in which the conditions were administered, no significant differences were observed.

4.1.3 Self-Report Measures

For the subjective stress measurement, the transition from pre- to post-TSST and f-TSST was assessed using the PANAS and the STADI. The t-test results, highlighting the differences between conditions, are detailed in Table 4.2. As illustrated in Figure 4.4, the average increase in the total PANAS score was significantly more pronounced in the TSST than in the f-TSST. Regarding the two subscales, the increase in Negative Affect following the TSST was higher compared to the f-TSST, while there was a larger increase during the f-TSST for Positive Affect.

Table 4.2: t-test results for the questionnaires PANAS and STADI.

Questionnaire	Subscale	t(43)	p	Hedges' g
PANAS	Negative Affect	2.550	0.152	0.618
	Positive Affect	-2.579	0.142	-0.646
	Total Score	-3.696	0.007**	-0.898
STADI	Anxiety	3.647	0.008**	0.955

In terms of the STADI, the Anxiety subscale showed a significantly greater increase from pre- to post-TSST than for f-TSST. No differences in effect sizes were observed for the condition order or the sitting/standing condition across any of the self-reported questionnaires.

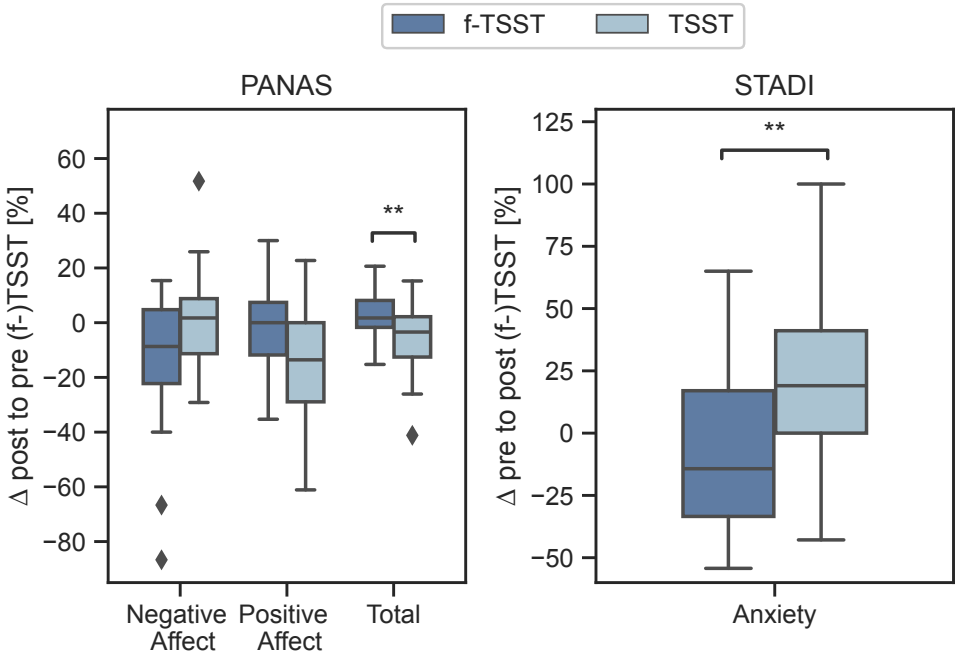

Figure 4.4: Questionnaire results for PANAS and STADI; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

4.2 Digital Biomarker

4.2.1 Inferential Statistics

Due to the large number of features, only the most relevant result will be shown in this section. The complete results of the statistical evaluation of all features can be found in Appendix B. Six participants were excluded due to missing video files or poor video quality. Thus, 37 participants were analyzed in this thesis (f-TSST: 37, TSST: 37).

Facial Expression Recognition

Emotions - Figure 4.5 displays the seven distinct emotions, detailing their mean values and standard deviation. Notably, happiness exhibited a significantly higher mean and standard deviation in the f-TSST compared to the TSST. Moreover, participants demonstrated a significantly higher frequency of neutral facial expressions in the TSST. Both conditions revealed the largest standard deviations for neutral expressions. Concerning other emotional expressions, the mean values for surprise and sadness were marginally lower, by only 2%, than those for happiness in the TSST.

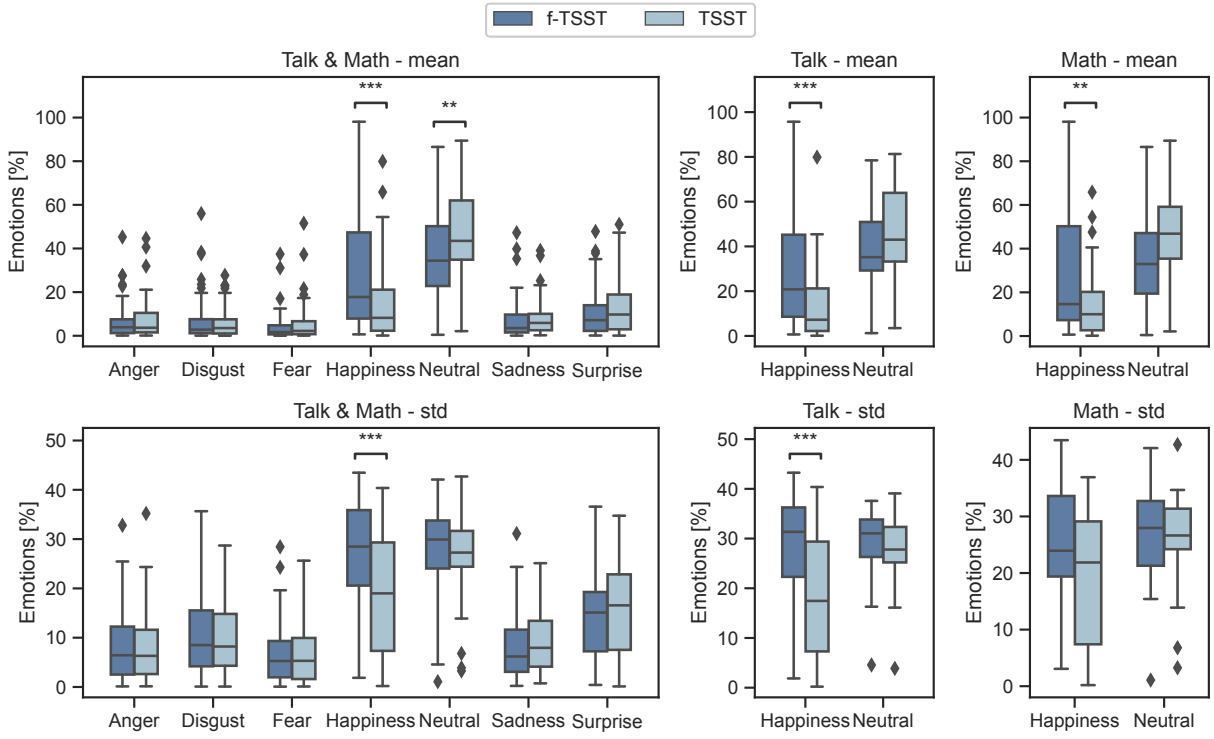

Figure 4.5: Mean and standard deviation of facial expressivity across all participants during the (f-)TSST, as well as each phase individually; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

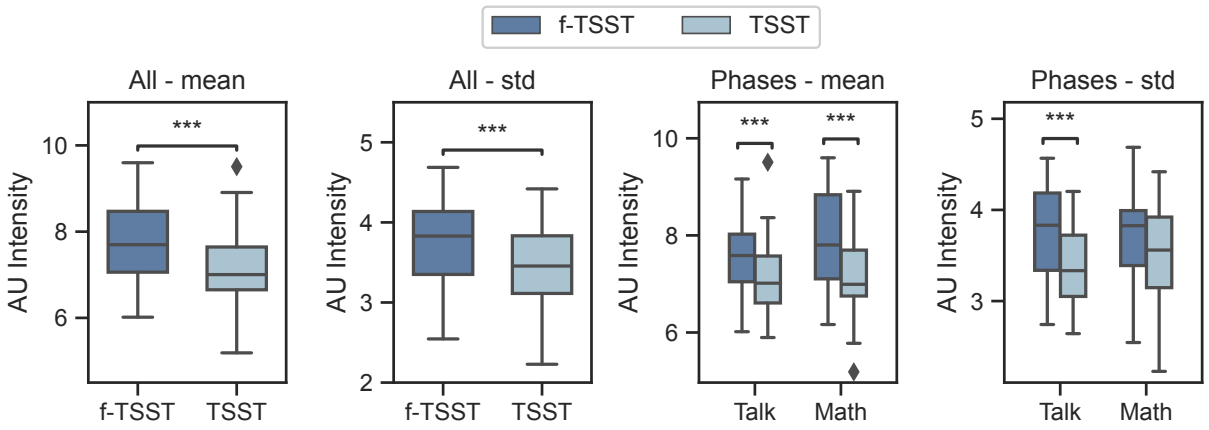

Figure 4.6: Mean and standard deviation results of AU intensity across all participants during the (f-)TSST, as well as each phase individually; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Upon examining the different phases, happiness was significantly elevated in both, with a more pronounced effect size during the *Talk* phase with a Hedges' g = -0.742. Although the standard deviation for happiness was greater in both phases, it was only significantly higher in the *Talk* phase. For neutral facial expressions, both phases recorded higher mean values in the TSST.

Action Units - Figure 4.6 demonstrates that the overall mean intensity of AUs was significantly higher in the f-TSST compared to the TSST. This includes not only the mean values but also the standard deviations, all of which were notably higher in the f-TSST.

The analysis identified specific AUs with significantly higher mean values in the f-TSST, as illustrated in Figure 4.7. These are: AU06 - Cheeck Raiser, AU09 - Nose Wrinkler, AU12 - Lip Corner Puller, AU14 - Dimpler, AU20 - Lip Stretcher, AU24 - Lips Part. Furthermore, the standard deviations for certain AUs were significantly greater, which include: AU06 - Cheek Raiser, AU07 - Lid Tightener, AU09 - Nose Wrinkler, AU10 - Upper Lip Raiser, AU12 - Lip Corner Puller.

When analyzing the different phases, both the mean and standard deviation of AU intensity showed an increase in the f-TSST across the *Math* and *Talk* phases, as indicated in Figure 4.6. For the individual AUs, all the AUs listed above, except for AU24, exhibited significantly higher mean values and standard deviations in both phases, as depicted in the Appendix A in Figure A.6.

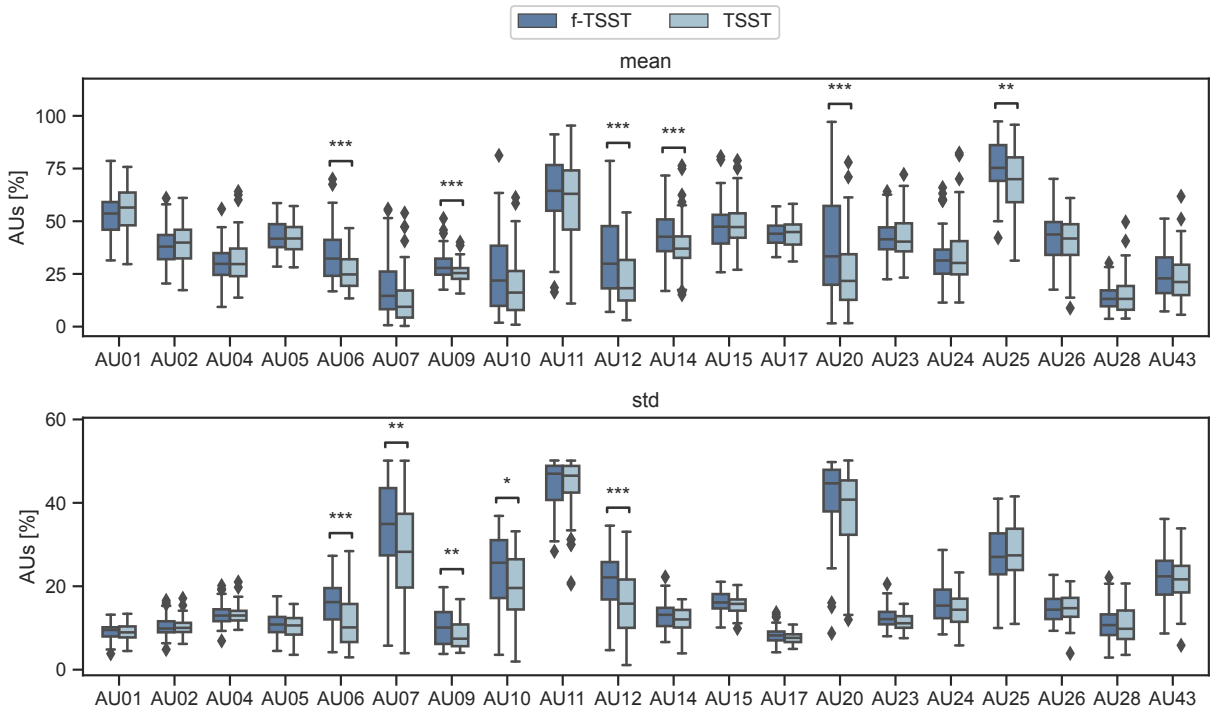

Figure 4.7: Mean and standard deviation results for all AUs across all participants; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Movement Patterns

The analysis focused on the movements of the head and both shoulders, revealing significant differences primarily associated with the head when comparing the TSST to the f-TSST. Results about shoulder movements are detailed in Figure A.7 in the Appendix A.

Generic Features - Figure 4.8 shows that the mean head velocity was lower during the TSST compared to the f-TSST. The same pattern was observed in the *Talk* and *Math* phases. Regarding the standard deviation of the head velocity and range of head motion, no significant differences were observed. The same holds for the visibility of the left and right hand as well as the elbow. Additional results are shown in the Appendix A in Figure A.7 for the generic features of the left and right shoulder, as well as in Figure A.8 for the features relating to the visibility of both hands and elbows.

Expert Features - An examination of the different upper body parts indicated that the percentage of movements falling below a specified threshold was elevated for the head in the TSST compared to

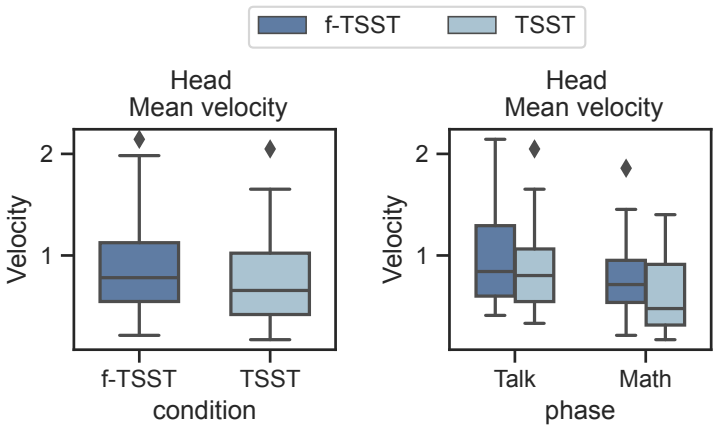

Figure 4.8: Results of the generic head movement features during the (f-)TSST across all participants, as well as for the individual phases; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

the f-TSST as highlighted in Figure 4.8. A detailed analysis of the phases, visualized in Figure 4.9, revealed that this increase was significant only during the *Math* phase. While the percentage of movements below the threshold for the head was also higher in the TSST during the *Talk* phase, this difference did not reach statistical significance.

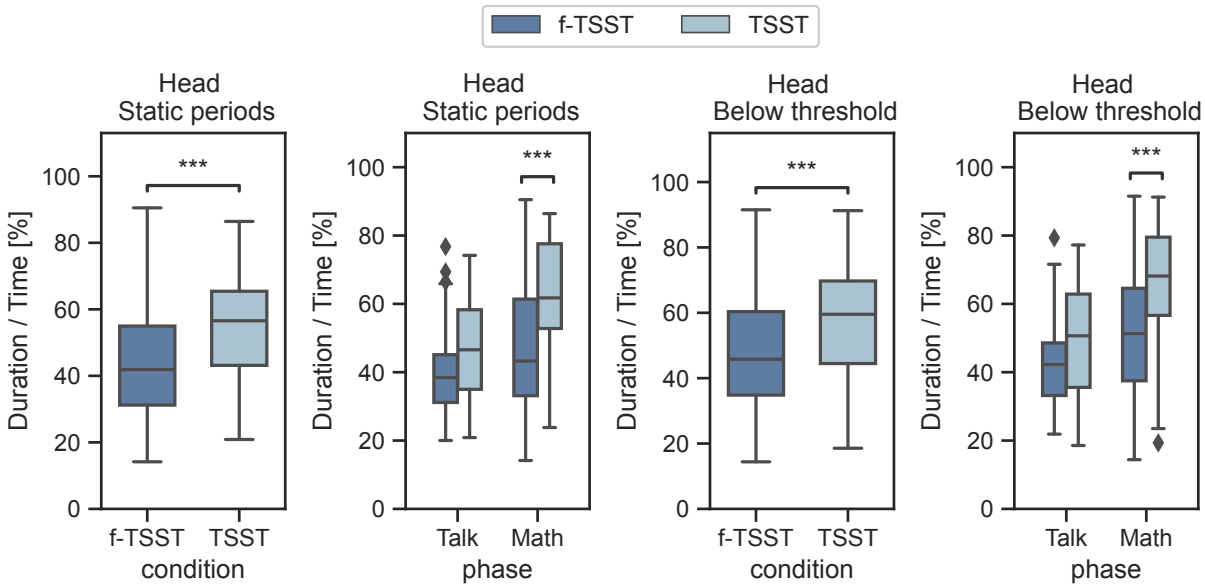

Figure 4.9: Results of the expert head movement features during the (f-)TSST across all participants, as well as for the individual phases; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

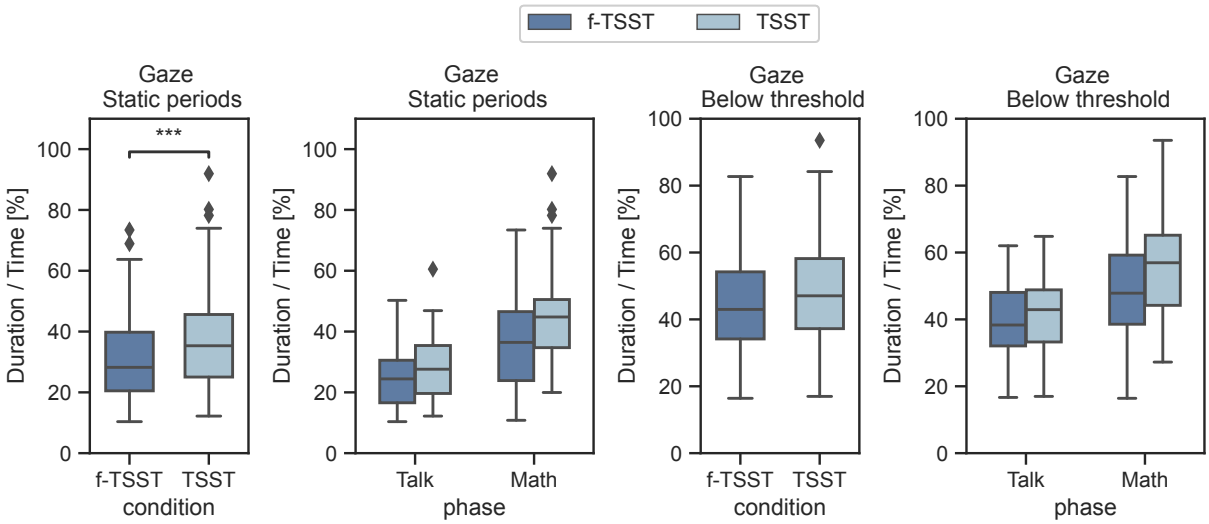

Figure 4.10: Gaze behavior results for static periods and gaze velocity falling below threshold across all participants during the (f-)TSST, as well as each phase individually; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Furthermore, a comparison of static periods of the head demonstrated a significantly higher occurrence in the TSST across both phases, shown in Figure 4.9. This increase was only significant during the *Math* phase.

Gaze Behavior

No significant differences were observed between the two conditions across all generic gaze velocity measures. However, static periods were significantly more frequent in the TSST compared to the f-TSST as shown in Figure 4.10. A more detailed analysis of the different phases revealed more static gaze behavior in the *Math* than in the *Talk* phase. For the measure of gaze velocity falling below threshold, a consistent pattern emerged, showing an overall higher percentage of gaze velocity below this threshold in the TSST, as well as in the *Math* phase when compared to the *Talk* phase.

Furthermore, analyses concerning pupil diameter revealed no significant differences between the TSST and f-TSST, as shown in Figure A.9 in the Appendix A. The number of blinks per minute similarly did not demonstrate any notable differences between the conditions.

4.2.2 ML-based Classification

Talk and Math

The classification of stress versus no-stress from digital biomarkers achieved a mean accuracy of 73.7% ± 8.2%, as detailed in Table 4.3. This accuracy was obtained with the following configuration:

- Scaler: Standard Scaler
- Feature Selection: RFE
- Classifier: AdaBoost

In Figure 4.11 the confusion matrix of the best-performing pipeline is displayed. Further analysis of the pipeline's performance across different condition orders revealed that the accuracy was higher when the TSST protocol was administered first (75.7% compared to 71.1%). The confusion matrices for each condition order are depicted in Figure 4.11.

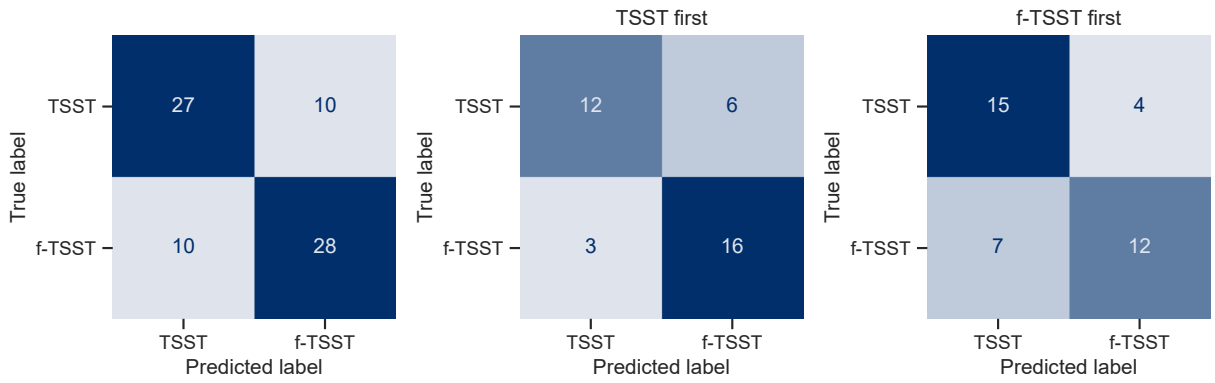

Figure 4.11: Confusion matrix results of best-performing classification pipeline trained on facial expression, body movement, and gaze features computed over the (f-)TSST. The two confusion matrices on the right side show the predictions for the two different condition orders.

A noteworthy difference in accuracy was observed between male and female conditions: female participants were classified with an accuracy of 83.3%, whereas for male participants an accuracy of 60.6% was achieved. Regarding sitting and standing conditions, no significant differences in accuracy were observed between classifying solely sitting participants (72.2%) and standing participants (74.4%)

SHAP values indicated that the most influential features in the best-performing classification pipeline, which was trained using the (f-)TSST data, included a combination of facial expressions, movements, and pupil dynamics. These findings are illustrated in Figure 4.12.

Table 4.3: Mean± standard deviation of classification performance metrics over the 5-fold model evaluation CV with features computed over the (f-)TSST. For each evaluated classifier, the classification pipeline combination with the highest mean accuracy is shown. The classification pipelines scoring the highest metrics are highlighted in **bold**.

Scaler	Feature Selection	Classifier	Accuracy [%]	F1-score [%]	Precision [%]
Standard	RFE	Ada	**73.7 (8.2)**	**72.2 (11.3)**	**73.2 (5.6)**
	SkB	NB	69.6 (12.2)	70.7 (14.2)	65.0 (10.0)
Min-Max	SkB	kNN	69.6 (5.4)	69.6 (9.9)	67.6 (4.5)
Standard	SkB	SVM	69.5 (6.0)	69.2 (10.2)	67.0 (4.8)
	RFE	RF	68.2 (9.2)	66.9 (10.7)	69.1 (9.1)
Min-Max	SkB	DT	64.3 (12.5)	60.0 (16.2)	70.7 (20.1)

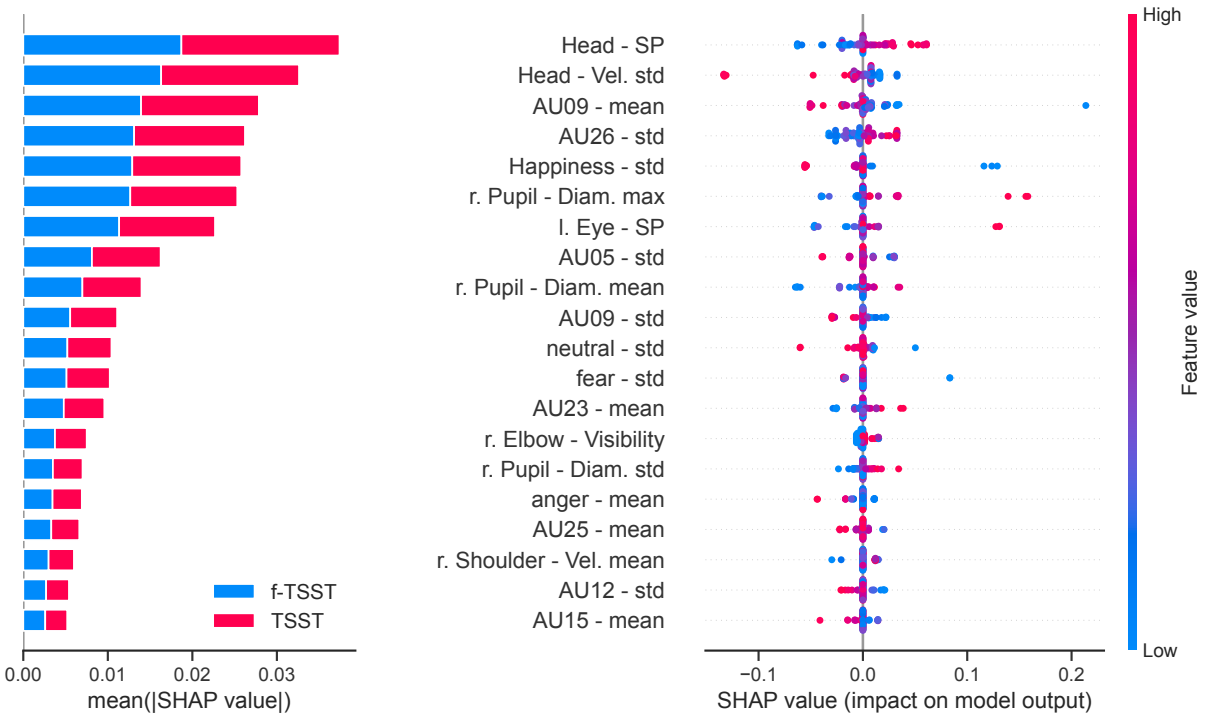

Figure 4.12: SHAP values computed over the whole (f-)TSST: Feature importances were determined using SHAP values calculated across the model evaluation cross-validation folds. Positive SHAP values correlate with an increased likelihood of predicting the positive class, namely TSST. Note: SP = Static Periods, r.=right, l.=left.

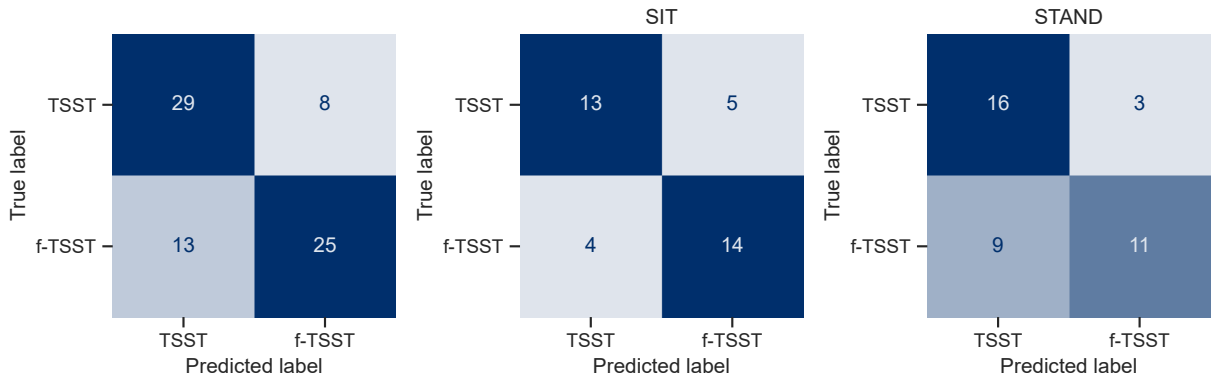

Figure 4.13: Confusion matrix results of best-performing classification pipeline trained on facial expression, body movement, and gaze features computed over the *Talk* phase during the (f-)TSST. The two confusion matrices on the right side show the predictions for the conditions sitting and standing.

Talk

When performing the same classification experiments using features only extracted during the Talk phase, the highest performing classifier reached an accuracy of 72.3%±12.7%, as presented in Table 4.4. A higher accuracy was observed when participants were sitting (75.0%) over standing (69.2%). The confusion matrices are shown in Figure 4.13. Furthermore, classifying solely female participants resulted in a higher accuracy (78.6%) than classifying male participants (63.6%). The analysis did reveal small differences in classification performance when dividing the data by condition order (TSST-first: 73.0% vs. f-TSST-first: 71.1%).

Table 4.4: Mean± standard deviation of classification performance metrics over the 5-fold model evaluation CV with features computed over the *Talk* phase during (f-)TSST. For each evaluated classifier, the classification pipeline combination with the highest mean accuracy is shown. The classification pipelines scoring the highest metrics are highlighted in **bold**.

Scaler	Feature Selection	Classifier	Accuracy [%]	F1-score [%]	Precision [%]
Min-Max	SkB	RF	**72.3 (12.7)**	**73.2 (13.2)**	69.5 (12.4)
	RFE	Ada	71.0 (11.7)	71.2 (11.9)	**70.3 (13.3)**
	SkB	SVM	69.6 (11.1)	72.5 (10.0)	66.4 (11.6)
		kNN	69.5 (7.1)	71.3 (8.0)	67.0 (9.4)
Standard	RFE	NB	67.0 (11.9)	64.4 (15.8)	66.9 (12.2)
		DT	63.9 (8.2)	63.7 (11.2)	63.8 (10.1)

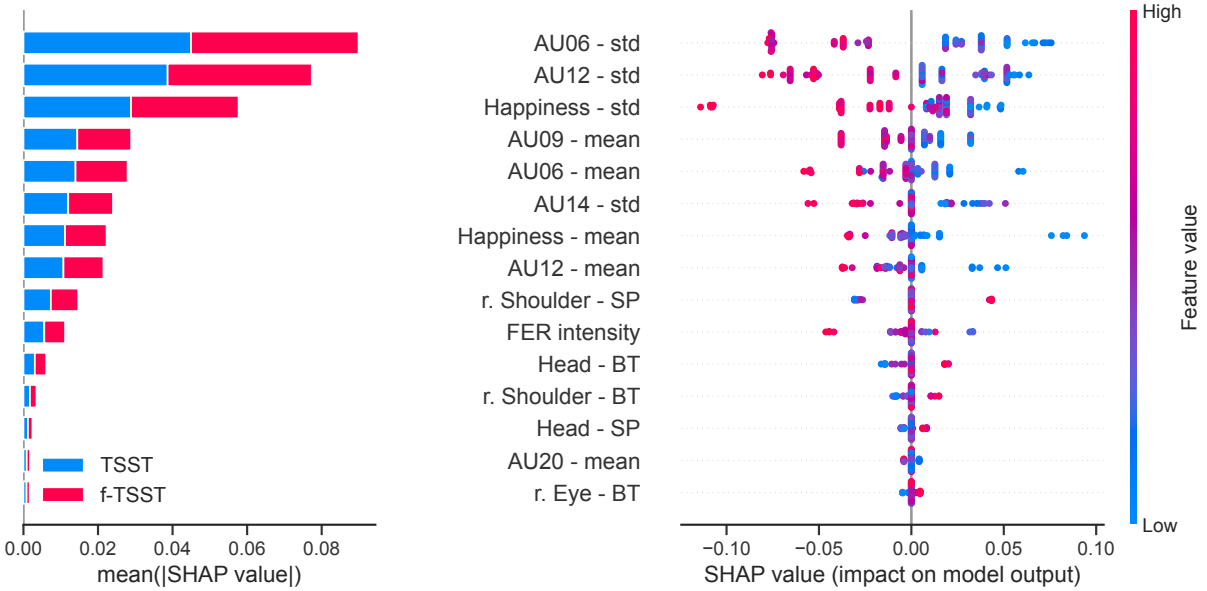

Figure 4.14: SHAP values computed over the *Talk* phase: Feature importances were determined using SHAP values calculated across the model evaluation cross-validation folds. A higher absolute SHAP value signifies greater influence on the model's classification output. Positive SHAP values correlate with an increased likelihood of predicting the positive class, namely TSST. Note: SP = Static Periods, BT = Below Threshold, r.=right, l.=left.

The resulting SHAP values of the best-performing pipeline for the *Talk* phase are shown in Figure 4.14. The explainable ML methods highlight that the most important features contributing most to the decision of the classier solely consist of facial expression features.

Math

In the *Math* phase, the classifier exhibited an accuracy of 73.5%±11.4%, as highlighted in Table 4.5. A slightly higher accuracy was achieved when the participants were standing than sitting (74.4% vs. 72.0%). Regarding gender differences, female participants were more likely correctly classified than male ones (Male: 63.6% vs. Female: 81.0%). The confusion matrices are shown in Figure 4.15. No major differences were found by dividing the data by condition order in terms of accuracy (TSST first: 73.0%, f-TSST first: 73.7%).

Figure 4.16 visualizes the most important features which had the biggest impact on the model, comprising of facial expressions, movement patterns, and pupil dynamics, with body movement features notably dominating the feature importance list.

Table 4.5: Mean± standard deviation of classification performance metrics over the 5-fold model evaluation CV with features computed over the *Math* phase during (f-)TSST. For each evaluated classifier, the classification pipeline combination with the highest mean accuracy is shown. The classification pipelines scoring the highest metrics are highlighted in **bold**.

Scaler	Feature Selection	Classifier	Accuracy [%]	F1-score [%]	Precision [%]
Min-Max	RFE	SVM	73.5(11.4)	74.9(10.9)	72.3(14.1)
		kNN	70.7 (10.2)	73.0 (8.7)	68.6 (10.8)
		DT	69.2 (7.4)	67.2 (7.8)	72.8 (10.6)
	SkB	RF	68.4 (8.0)	69.7 (9.9)	65.1 (4.9)
		NB	68.2 (5.6)	70.1 (7.7)	64.6 (3.5)
Standard	SkB	Ada	67.2 (12.1)	65.7 (17.2)	65.9 (12.4)

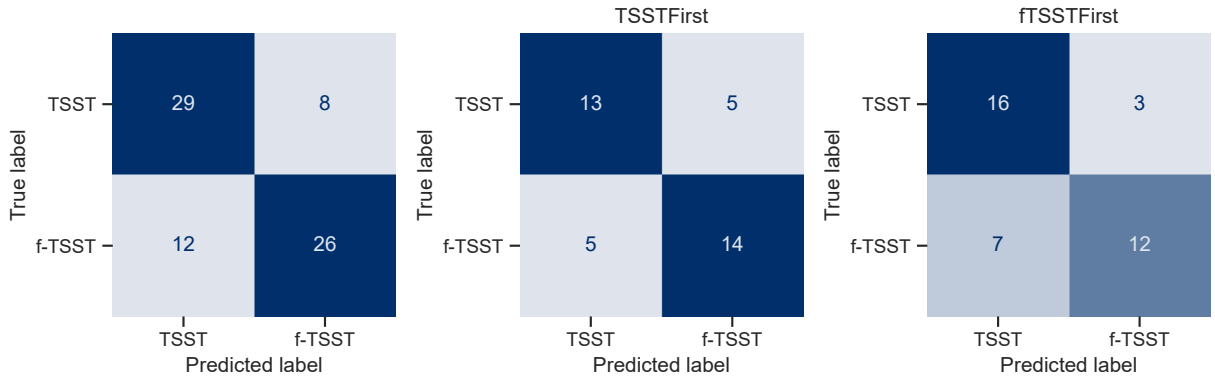

Figure 4.15: Confusion matrix results of best-performing classification pipeline trained on facial expression, body movement and gaze features computed over the *Math* phase during the (f-)TSST. The two confusion matrices on the right side show the predictions for the two different condition orders.

4.2.3 SBMLR

PCA components were calculated for the differences in facial expression, movement, and gaze features between TSST andf-TSST, each while maintaining an explained variance of 80%. For predicting m_{S1S4} the highest explained variance, 0.644, was achieved using 10 facial expression features, alongside two movement and three gaze components, as shown in Table 4.6. The total PANAS score was best predicted by five facial expression features, one movement, and two gaze components, achieving an explained variance of 0.540 (Table 4.7). Additionally, 0.584 of the variance in mean HR could be explained by a model derived from 10 principal components

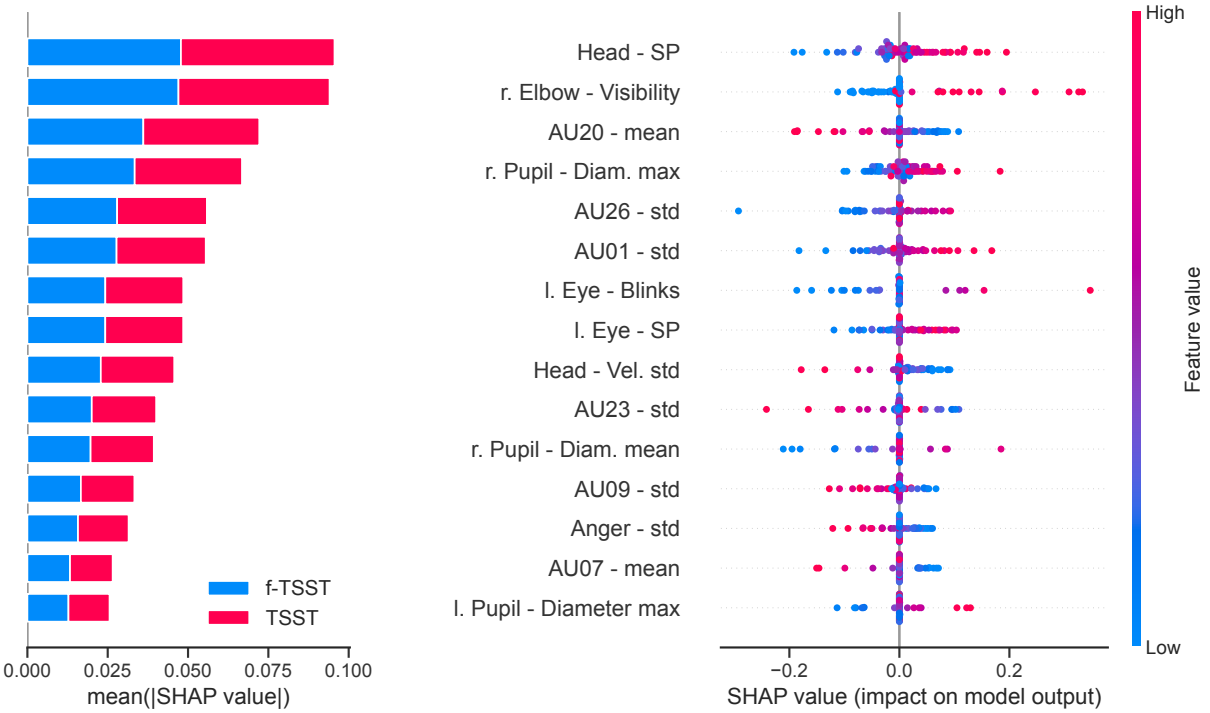

Figure 4.16: SHAP values computed over the *Math* phase: Feature importances were determined using SHAP values calculated across the model evaluation cross-validation folds. A higher absolute SHAP value signifies greater influence on the model's classification output. Positive SHAP values correlate with an increased likelihood of predicting the positive class, namely TSST.
Note: SP = Static Periods, BT: Below Threshold, r.=right, l.=left.

(Table 4.8).

During different phases, the prediction of m_{S1S4} and total PANAS score resulted in a similar explained variance for the *Math* phase, with 64.15% for slope and 74.47% for the total PANAS score. For the *Talk* phase, the explained variance was lower, at 48.73% for m_{S1S4} and 42.79% for the total PANAS score. Conversely, for mean HR in the *Talk* phase, the explained variance reached 60.27% using eight facial expression features, three movement, and one gaze feature. Predicting mean HR in the *Math* phase resulted in an explained variance of 38.5%.

Table 4.6: Results of linear regression predicting m_{S1S4} with features of facial expression, movement and gaze patterns; β: standardized regression coefficient; σ: standard error; adj.: adjusted

components	β	σ	t	p	R^2	adj. R^2	CI[2.5%]	CI[97.5%]
Fer_PCA_2	0.028	0.016	1.728	0.102	0.811	0.644	-0.006	0.061
Fer_PCA_4	0.033	0.026	1.255	0.227	0.811	0.644	-0.022	0.087
Fer_PCA_5	-0.067	0.027	-2.493	0.023	0.811	0.644	-0.123	-0.010
Fer_PCA_6	0.035	0.025	1.376	0.187	0.811	0.644	-0.019	0.088
Fer_PCA_7	-0.043	0.023	-1.858	0.081	0.811	0.644	-0.092	0.006
Fer_PCA_9	0.119	0.037	3.209	0.005	0.811	0.644	0.041	0.198
Fer_PCA_10	-0.111	0.034	-3.316	0.004	0.811	0.644	-0.182	-0.040
Fer_PCA_11	0.036	0.032	1.129	0.275	0.811	0.644	-0.031	0.103
Fer_PCA_12	0.181	0.035	5.118	0.000	0.811	0.644	0.106	0.255
Fer_PCA_13	-0.136	0.038	-3.584	0.002	0.811	0.644	-0.216	-0.056
Mov_PCA_2	0.163	0.034	4.758	0.000	0.811	0.644	0.091	0.235
Mov_PCA_3	-0.147	0.045	-3.276	0.004	0.811	0.644	-0.241	-0.052
Eyes_PCA_0	-0.047	0.026	-1.846	0.082	0.811	0.644	-0.101	0.007
Eyes_PCA_1	0.141	0.028	4.987	0.000	0.811	0.644	0.081	0.201
Eyes_PCA_2	-0.159	0.050	-3.200	0.005	0.811	0.644	-0.263	-0.054

Table 4.7: Results of linear regression predicting PANAS total score with features of facial expression, movement and gaze patterns; β: standardized regression coefficient; σ: standard error; adj.: adjusted

components	β	σ	t	p	R^2	adj. R^2	CI[2.5%]	CI[97.5%]
Fer_PCA_0	0.074	0.020	3.727	0.001	0.659	0.540	0.033	0.115
Fer_PCA_1	0.057	0.025	2.282	0.032	0.659	0.540	0.005	0.108
Fer_PCA_9	-0.114	0.046	-2.494	0.020	0.659	0.540	-0.208	-0.019
Fer_PCA_12	-0.085	0.053	-1.602	0.123	0.659	0.540	-0.195	0.025
Fer_PCA_13	0.128	0.060	2.123	0.045	0.659	0.540	0.003	0.252
Mov_PCA_3	0.261	0.074	3.510	0.002	0.659	0.540	0.107	0.415
Eyes_PCA_1	-0.224	0.043	-5.270	0.000	0.659	0.540	-0.312	-0.136
Eyes_PCA_2	0.096	0.069	1.397	0.176	0.659	0.540	-0.046	0.239

Table 4.8: Results of linear regression predicting mean HR with features of facial expression, movement and gaze patterns; β: standardized regression coefficient; σ: standard error; adj.: adjusted

components	β	σ	t	p	R^2	adj. R^2	CI[2.5%]	CI[97.5%]
Fer_PCA_0	0.712	0.570	1.250	0.223	0.706	0.584	-0.464	1.889
Fer_PCA_1	3.172	0.764	4.153	0.000	0.706	0.584	1.596	4.748
Fer_PCA_3	-1.364	0.730	-1.869	0.074	0.706	0.584	-2.870	0.142
Fer_PCA_4	1.616	1.026	1.575	0.128	0.706	0.584	-0.501	3.734
Fer_PCA_6	3.529	1.153	3.061	0.005	0.706	0.584	1.150	5.909
Fer_PCA_7	-2.437	1.181	-2.064	0.050	0.706	0.584	-4.874	-0.001
Fer_PCA_8	-6.536	1.314	-4.976	0.000	0.706	0.584	-9.247	-3.825
Mov_PCA_2	7.100	1.449	4.900	0.000	0.706	0.584	4.109	10.090
Mov_PCA_3	4.783	2.253	2.123	0.044	0.706	0.584	0.132	9.434
Eyes_PCA_2	7.604	2.174	3.498	0.002	0.706	0.584	3.118	12.091

4.2.4 ML-based Regression

ML techniques applied to predict Positive and Negative Affect Scores of the PANAS showed negative R^2 values for both positive and negative emotions. Attempts to model mean HR also yielded negative R^2 values. Further analysis during the *Talk* and *Math* phases of the test continued to reveal negative R^2 values for both PANAS scores and mean HR.

For the prediction of m_{S1S4}, the results were modestly better, but still limited. The best performing model, employing StandardScaler for data normalization, RFE for feature selection, and KNeighborsRegressor for prediction, achieved a R^2 of 0.02$\pm$0.15 and a MAE of 0.130$\pm$0.12 nmol/L. No significant improvements in prediction accuracy were observed by either optimizing for the MAE or by analyzing data from the *Math* and *Talk* phases separately.

4.3 rPPG

4.3.1 Validation

For the conventional methods, POS performed the best on UBFC-rPPG with a MAE of 3.22 bpm. On PURE, LGI achieved the lowest MAE with 5.07 bpm, as well as on COHFACE with a MAE of 21.07 bpm. The results for PURE and UBFC-rPPG are shown in Table 4.9 and for COHFACE in Table 4.10.

Table 4.9: rPPG performance of conventional models on UBFC-rPPG and PURE. The best results are highlighted. Note: MAE and RMSE are expressed in beats per minute, while MAPE is expressed as a percentage (%).

	UBFC-rPPG				**PURE**			
Model	MAE	MAPE	ρ	RMSE	MAE	MAPE	ρ	RMSE
ICA	11.33	10.58	0.51	21.61	5.56	5.42	0.68	17.52
GREEN	14.48	13.56	0.41	26.36	10.01	10.82	0.35	23.17
CHROM	6.07	6.09	0.82	10.88	7.81	15.10	0.69	18.54
LGI	9.94	9.29	0.53	21.28	**5.07**	**5.22**	0.72	16.59
POS	**3.22**	**3.02**	**0.85**	**9.79**	5.40	10.18	**0.78**	**15.30**

Table 4.10: rPPG performance of conventional models on COHFACE. The best results are highlighted. Note: MAE and RMSE are expressed in beats per minute, while MAPE is expressed as a percentage (%).

	COHFACE			
Model	MAE	MAPE	ρ	RMSE
ICA	26.70	54.46	-0.06	30.99
GREEN	29.19	59.60	-0.11	33.74
CHROM	29.76	60.57	-0.05	32.65
LGI	**21.07**	**43.08**	**-0.04**	**25.28**
POS	29.47	60.12	-0.08	32.41

Regarding deep learning-based rPPG, models trained on PURE and UBFC-rPPG datasets were cross-tested against each other. The best performance was observed when models were trained on PURE and tested on UBFC-rPPG, where TSCAN achieved a MAE of 4.50 bpm. Conversely, when models were trained on UBFC-rPPG and tested on PURE, TSCAN also was the best performer with a MAE of 6.96 bpm. For COHFACE, the lowest MAE with 17.39 bpm was achieved by DeepPhys when trained on PURE. These findings are summarized in Table 4.11. Notably, the two

Table 4.11: rPPG performance of deep learning models across datasets for cross-testing and COHFACE evaluation. For cross-testing, the results on UBFC-rPPG (models trained on PURE) and on PURE (models trained on UBFC-rPPG) are shown. The best performing models are highlighted. Note: MAE and RMSE are expressed in beats per minute, while MAPE is expressed as a percentage (%).

		Cross-testing				**COHFACE**			
Trained	Model	MAE	MAPE	ρ	RMSE	MAE	MAPE	ρ	RMSE
PURE	DeepPhys	7.39	7.56	0.56	14.98	**17.39**	**35.12**	0.02	**20.46**
	EfficientPhys	6.73	6.95	0.55	14.61	32.28	65.55	0.06	36.56
	PhysFormer	6.59	6.36	0.68	13.24	18.91	37.77	0.00	24.94
	PhysNet	10.57	10.70	0.45	16.92	19.39	38.45	0.17	24.94
	TSCAN	**4.50**	**4.32**	**0.73**	**11.49**	18.63	37.75	-0.04	21.51
UBFC-rPPG	DeepPhys	7.78	8.51	0.47	19.83	32.43	65.69	0.09	37.36
	EfficientPhys	7.54	8.28	0.52	18.88	37.62	76.19	0.10	41.14
	PhysFormer	13.95	21.18	0.21	23.95	34.57	68.72	0.03	37.21
	PhysNet	9.56	16.73	**0.67**	**18.03**	32.00	63.14	**0.18**	38.07
	TSCAN	**6.96**	**7.94**	0.58	18.22	19.34	39.23	-0.09	22.55

conventional methods POS and LGI both performed better on PURE and UBFC-rPPG than any of the deep learning network.

Figure 4.19 presents the predictions of TSCAN trained on PURE for the UBFC-rPPG, COHFACE, and UBFC-PHYS datasets, while for the PURE dataset the predictions of TSCAN trained on UBFC-rPPG are illustrated. For each dataset, predictions from multiple participants over durations ranging from four to eight minutes are depicted to provide a comprehensive overview. Notably, for PURE, UBFC-rPPG, and UBFC-PHYS, the predictions exhibit close alignment with the ground truth. However, for COHFACE, the mean ground truth HR is approximately 20 bpm lower than the predicted HR.

4.3.2 Comparing EmpkinS-TSST with UBFC-PHYS

For the datasets which include the (f-)TSST, deep learning as well as conventional models performed better on the UBFC-PHYS than on the EmpkinS-TSST. Among these, PhysNet, trained on PURE, achieved the lowest MAE at 8.10 bpm for UBFC-PHYS. POS performed the best compared to the other conventional methods with a MAE of 8.82 bpm. It's noteworthy that the performance of the different deep learning models was relatively consistent, with MAEs ranging from 8.10 bpm to 10.08 bpm. In contrast, for the EmpkinS-TSST, the two best-performing models were TSCAN

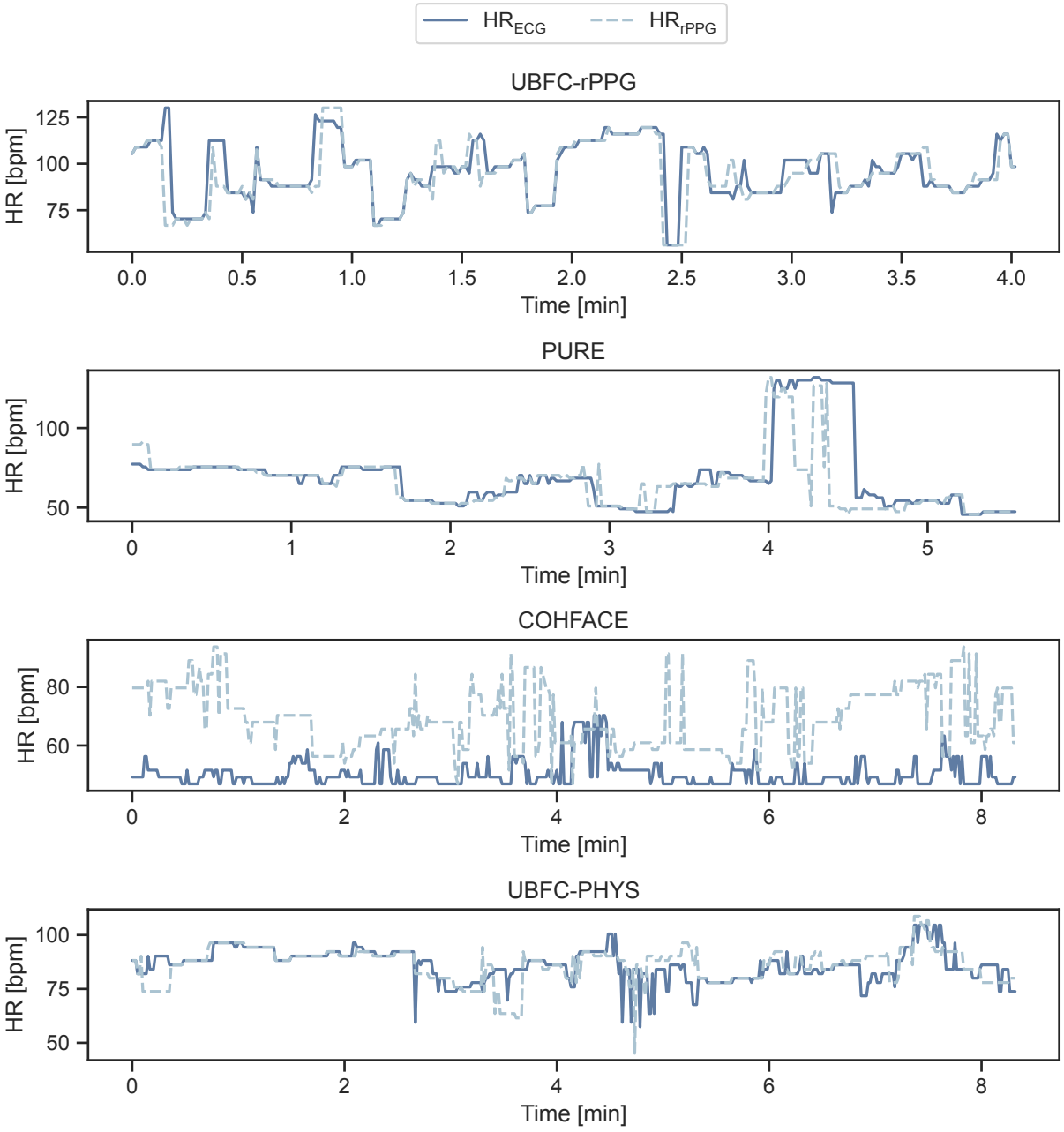

Figure 4.17: Selected HR predictions of TSCAN trained on PURE for the datasets UBFC-rPPG, COHFACE and UBFC-PHYS. For PURE, the HR predictions from TSCAN trained on UBFC-rPPG are shown.

Table 4.12: rPPG performance of conventional models on UBFC-PHYS and EmpkinS-TSST. The best results are highlighted. Note: MAE and RMSE are expressed in beats per minute, while MAPE is expressed as a percentage (%).

	UBFC-PHYS				**EmpkinS-TSST**			
Model	MAE	MAPE	ρ	RMSE	MAE	MAPE	ρ	RMSE
ICA	10.15	13.01	0.35	14.89	29.03	27.59	0.01	36.24
GREEN	15.51	18.42	0.12	20.17	32.91	31.38	0.02	39.67
CHROM	10.40	13.92	0.38	14.37	**21.24**	**20.31**	**0.11**	**28.22**
LGI	10.31	13.28	0.34	16.15	25.94	24.50	0.07	33.84
POS	**8.82**	**12.02**	**0.43**	**14.06**	21.40	20.63	0.07	28.32

Table 4.13: rPPG performance of deep learning models on UBFC-PHYS and Empkins-TSST. The best results are highlighted. Note: MAE and RMSE are expressed in beats per minute, while MAPE is expressed as a percentage (%).

		UBFC-PHYS				**EmpkinS-TSST**			
Trained	Model	MAE	MAPE	ρ	RMSE	MAE	MAPE	ρ	RMSE
PURE	DeepPhys	10.08	12.79	0.40	15.67	22.32	21.59	0.13	29.90
	EfficientPhys	8.64	11.40	0.45	14.24	18.30	18.14	0.20	25.44
	PhysFormer	8.41	11.31	**0.47**	13.95	26.64	31.17	0.13	34.22
	PhysNet	**8.10**	**10.66**	0.46	**13.42**	21.40	22.86	0.14	28.41
	TSCAN	9.11	12.06	0.42	14.53	19.63	19.63	0.13	26.34
UBFC -rPPG	DeepPhys	9.48	12.08	0.42	15.03	21.06	20.32	0.19	28.77
	EfficientPhys	8.70	11.55	0.44	14.15	17.90	17.89	**0.25**	24.78
	PhysFormer	9.69	13.33	0.41	15.45	**16.54**	17.58	0.23	**22.75**
	PhysNet	8.75	12.12	0.45	14.65	18.36	20.69	0.19	24.88
	TSCAN	8.47	11.27	**0.47**	13.91	17.15	**17.22**	**0.25**	24.31

and PhysFormer, both trained on UBFC-rPPG. PhysFormer exhibited the lowest MAE at 16.54 bpm and the lowest RMSE at 22.75 bpm, while TSCAN recorded the lowest MAPE at 17.22% and the highest ρ at 0.25. Regarding the conventional models, CHROM achieved the lowest MAE with 21.24 bpm. The results are summarized in Table 4.12 for the conventional ones in Table 4.13 for the deep learning models

When comparing the different phases — *Pause*, *Talk*, and *Math* — for the best performing deep learning models the lowest MAE was observed during the *Pause* phase, and the highest during the *Talk* phase across both datasets. Furthermore, the MAE was consistently lower in the UBFC-PHYS than in the EmpkinS-TSST. Overall, the MAE was smaller in the f-TSST compared

to the TSST. The predicted HR for each phase is detailed in Table 4.14b. Remarkably, the mean rPPG-HR for the f-TSST was higher than that of the TSST data from UBFC-PHYS for all phases.

The output of the best-performing model for both the EmpkinS-TSST and UBFC-PHYS datasets was analyzed using a Bland-Altman plot, as illustrated in Figure 4.18. The mean difference in HR for the UBFC-PHYS dataset was -1.74 bpm, lower than that for the EmpkinS-TSST, which was 8.15 bpm. The EmpkinS-TSST exhibited greater variability in the differences between HR and rPPG-HR compared to UBFC-PHYS, as indicated by its wider limits of agreement. The range of mean HR and rPPG-HR spanned from 50 to 150 bpm for EmpkinS-TSST, in contrast to 50 to 125 bpm for UBFC-PHYS.

Table 4.14: rPPG model performance and predicted mean HR per phase for the EmpkinS-TSST (TSCAN trained on UBFC-rPPGS) and UBFC-PHYS (PhysNet trained on PURE). Note: MAE and mean HR are expressed in beats per minute.

(a) MAE results per phase.

	UBFC-PHYS	**EmpkinS-TSST**	
Phase	TSST	TSST	f-TSST
Pause	4.82	14.68	11.4
Talk	11.67	21.42	17.04
Math	8.18	18.43	15.28

(b) Predicted HR.

	UBFC-PHYS	**EmpkinS-TSST**	
Phase	TSST	TSST	f-TSST
Pause	81.34	95.65	86.22
Talk	80.79	103.46	94.37
Math	82.17	102.72	91.46

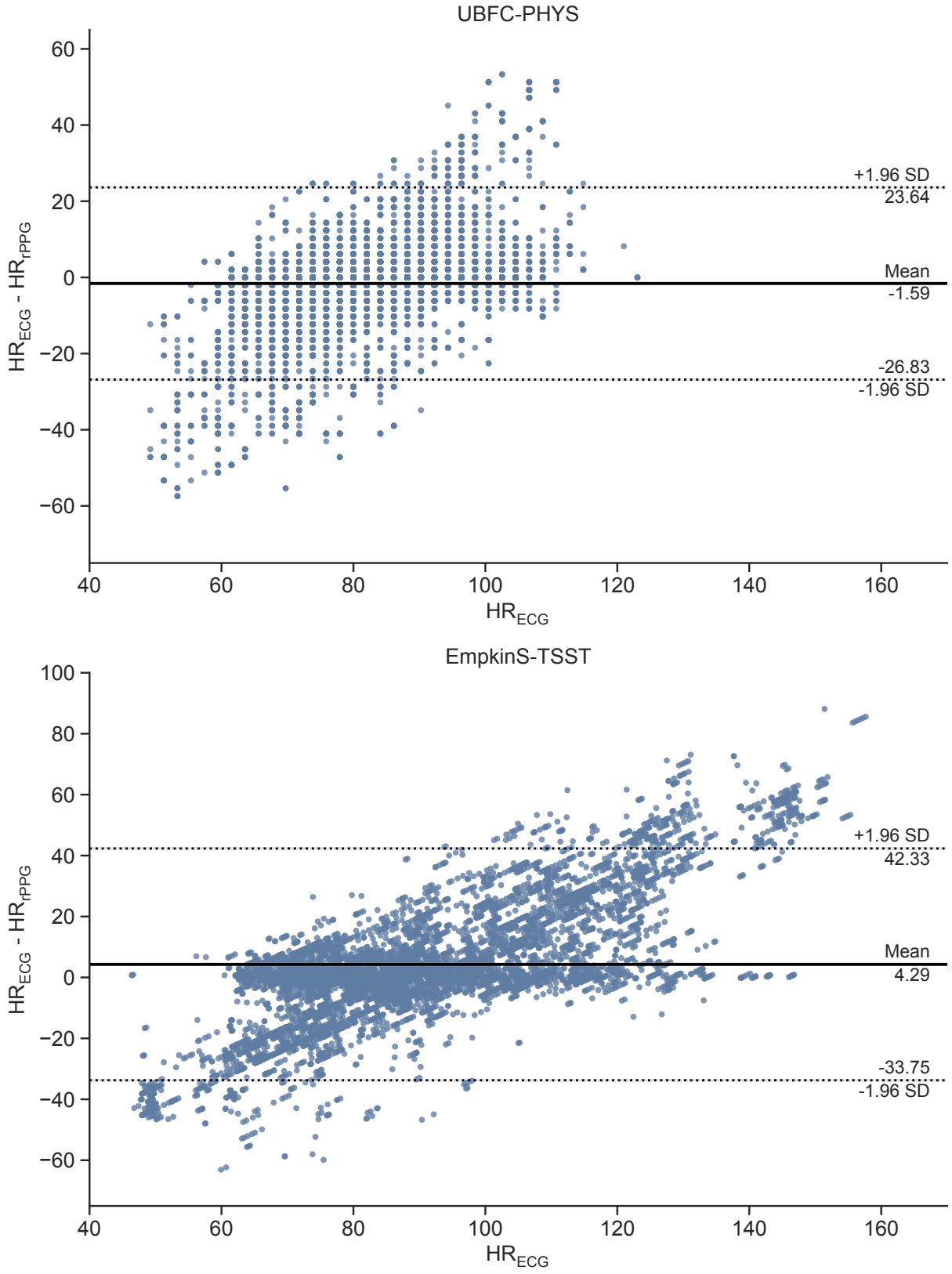

Figure 4.18: Bland-Altman Plots of the best-performing models for both the UBFC-PHYS and EmpkinS-TSST dataset.

4.4 Multimodal Stress State Detection

For the multimodal stress state detection, the predicted HR from the best performing rPPG model was used, namely TSCAN trained on UBFC-rPPG. In the following section, it will be referred to as rPPG-HR.

4.4.1 Inferential Statistics

As shown in Figure 4.19, the ECG-based HR was significantly higher in the TSST than in the f-TSST in all three phases *Pause*, *Talk* and *Math*. When taking a look at the predicted rPPG-HR, the rPPG-HR was significantly higher in the phases *Pause* and *Math*. Regarding the *Talk* phase, no differences were found. Figure 4.20 also shows, that the predicted rPPG-HR aligns more with the ECG-based HR in the *Pause* and *Math* phases than in the *Talk* phase for both f-TSST with lower HRs and TSST with higher HRs. For the *Talk* phase, the prediction aligns less with the the ground truth, which was also highlighted with higher MAEs in Table 4.14a in Section 4.3.2.

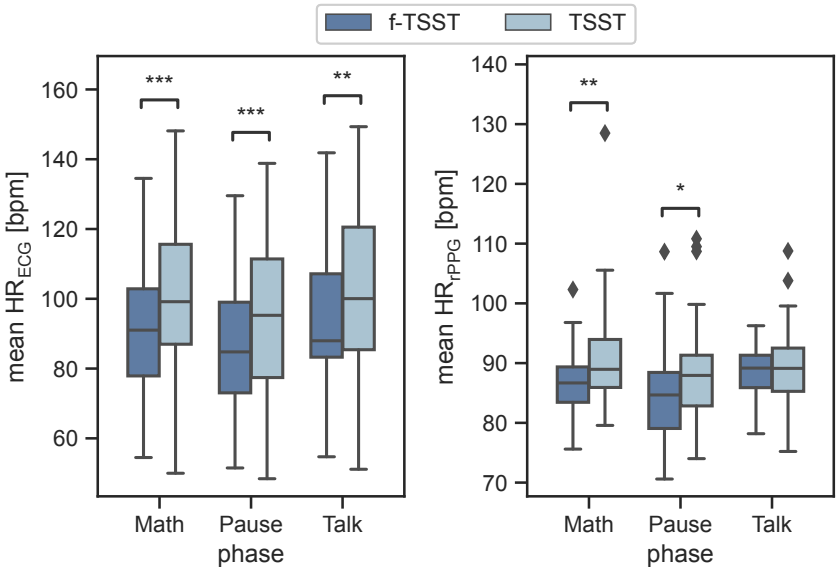

Figure 4.19: Mean HR distribution over the different phases *Pause*, *Talk*, and *Math*. On the left side the ground truth ECG-based HR is depicted and on the right side the predicted rPPG-HR; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

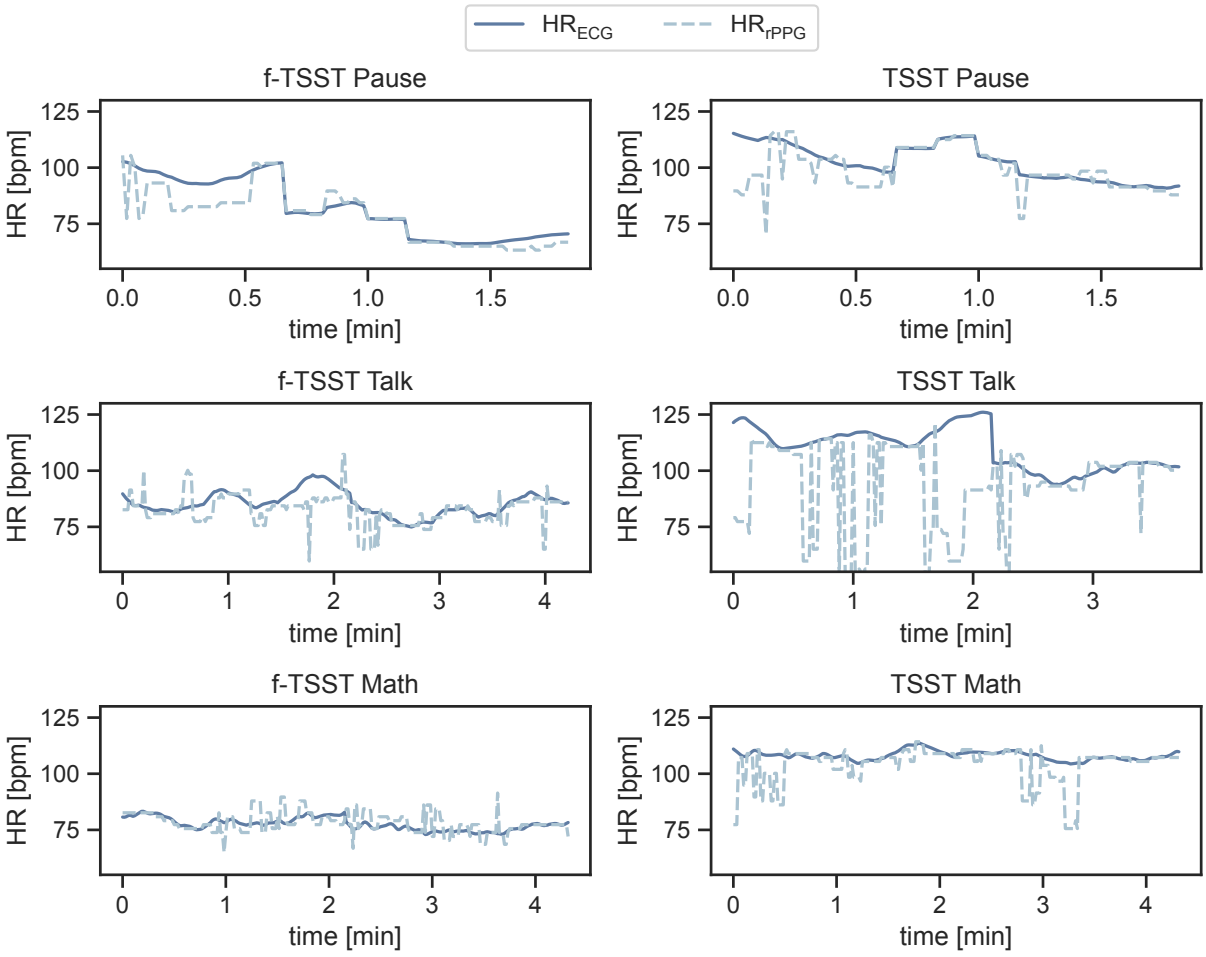

Figure 4.20: Predicted rPPG-HR compared to the ground truth ECG-based HR for one participant for the whole (f-)TSST over all three phases *Pause*, *Talk*, and *Math*.

4.4.2 ML-based Classification

Math and Talk

The classification of stress versus no-stress, integrating digital biomarkers with rPPG-HR, achieved a mean accuracy of 73.3%±5.5%, as outlined in Table 4.15. The resulting confusion matrix is shown in Figure 4.21. The accuracy was attained with the following configuration:

- Scaler: Standard Scaler
- Feature Selection: Select k best
- Classifier: kNN

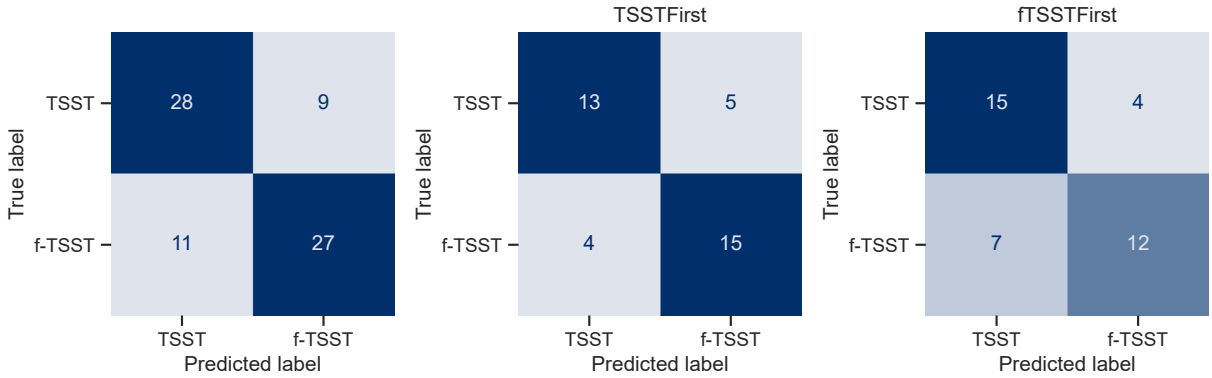

Figure 4.21: Confusion matrix results of best-performing classification pipeline trained on facial expression, body movement and gaze features computed over the (f-)TSST. The two confusion matrices on the right side show the predictions for the two different condition orders.

When classifying solely sitting participants, a higher accuracy was achieved (77.8%) than for sitting ones (69.2%). Regarding the participant's gender, a higher accuracy was observed for female participants (76.2%) than for male participants (69.7%). Further, a notable difference was observed in accuracy between different condition orders (TSST-First: 75.7%, f-TSST-First: 71.1%).

Regarding SHAP values, the top features consisted of facial expression, head movement, and gaze features, with mean rPPG-HR ranking at sixth position. All SHAP values are presented in Figure 4.22.

Table 4.15: Mean± standard deviation of classification performance metrics over the 5-fold model evaluation CV with multimodal features computed over the (f-)TSST. For each evaluated classifier, the classification pipeline combination with the highest mean accuracy is shown. The classification pipelines scoring the highest metrics are highlighted in **bold**.

Scaler	Feature Selection	Classifier	Accuracy [%]	F1-score [%]	Precision [%]
Standard	SkB	kNN	**73.3 (5.6)**	**72.4 (9.8)**	73.2 (7.3)
Min-Max	SkB	Ada	73.2 (9.3)	70.5 (16.6)	**74.3 (13.1)**
Standard	SkB	SVM	72.1 (6.1)	73.3 (6.7)	69.6 (5.1)
Min-Max	SkB	RF	69.8 (11.5)	67.1 (14.5)	69.8 (11.7)
Standard	SkB	NB	66.8 (13.3)	67.7 (16.5)	62.3 (11.9)
Min-Max	SkB	DT	65.7 (8.6)	65.0 (12.5)	64.8 (10.4)

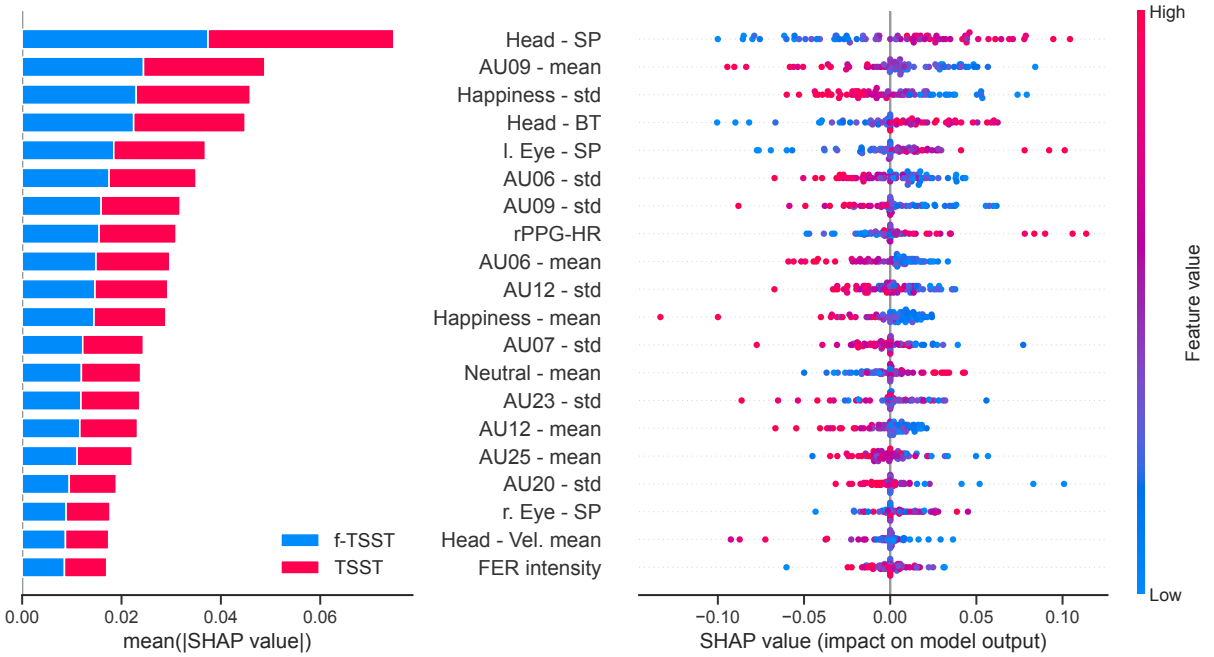

Figure 4.22: SHAP values of the multimodal approach for *Math* and *Talk*: Feature importances were determined using SHAP values calculated across the cross-validation folds of model evaluation. A higher absolute SHAP value signifies greater influence on the model's classification output. Positive SHAP values correlate with an increased likelihood of predicting the positive class, namely TSST. Note: SP = Static Periods.

Talk

In the *Talk* phase, training various classifiers with their respective hyperparameters on digital biomarkers combined with rPPG-HR identified the same optimal classifier as when the models were trained exclusively on digital biomarkers. Consequently, the detailed classification outcomes for the *Talk* phase are omitted here, as they align with those previously discussed and are illustrated in the Section 4.2.2 in Figure 4.13.

Math

Regarding the *Math* phase, the best-performing classifier exhibited an accuracy of 77.3%±6.5% and a F1-score of 79.3%±5.1%, employing Standard Scaler, SelectKBest for feature selection, and kNN. A higher accuracy was observed when participants conducted the f-TSST first (84.2%) than when the TSST was administered first (70.3%). Further, differences were found between female and male participants in terms of accuracy (Male: 81.8%, Female: 73.8%), as well as for the condition order (Sitting: 77.8%, Standing: 76.9%).

SHAP values revealed that mean rPPG-HR had the fourth largest impact on classification accuracy. The five most influential features included a mix of AUs and head movement features. Overall, the top 20 features span emotional values, AUs, head movements, and gaze features, as detailed in Figure 4.24.

Table 4.16: Mean± standard deviation of classification performance metrics over the 5-fold model evaluation CV with multimodal features computed over the *Math* phase during the (f-)TSST. For each evaluated classifier, the classification pipeline combination with the highest mean accuracy is shown. The classification pipelines scoring the highest metrics are highlighted in **bold**.

Scaler	Feature Selection	Classifier	Accuracy [%]	F1-score [%]	Precision [%]
Standard	SkB	kNN	**77.3 (6.5)**	**79.3 (5.1)**	**74.3 (9.5)**
Min-Max	SkB	SVM	70.7 (8.6)	72.8 (7.0)	68.7 (9.3)
	RFE	RF	69.7 (11.2)	68.6 (12.1)	71.3 (15.0)
	SkB	NB	69.3 (3.1)	71.5 (2.9)	66.6 (4.4)
Standard	RFE	Ada	67.8 (8.4)	66.2 (8.3)	72.8 (16.3)
Min-Max	RFE	DT	64.0 (5.3)	64.7 (4.6)	63.9 (8.2)

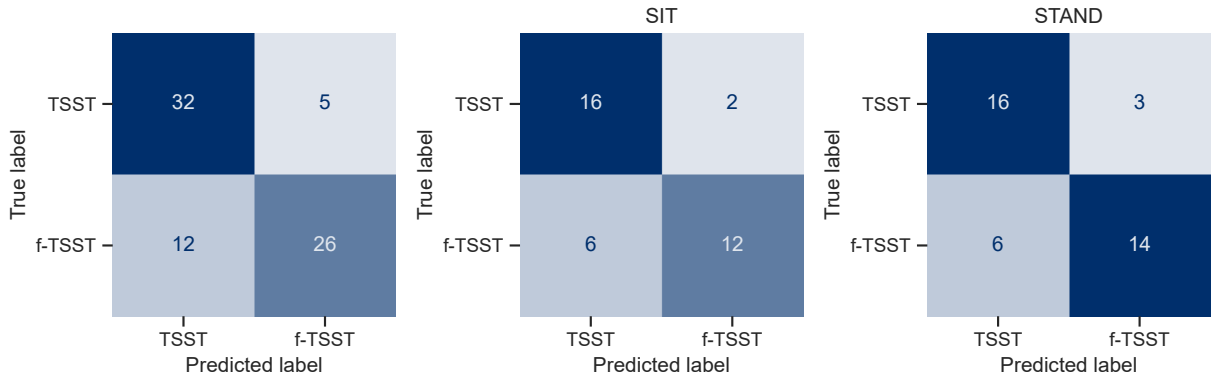

Figure 4.23: Confusion matrix results of best-performing classification pipeline trained on facial expression, body movement and gaze features computed over the *Math* phase during the (f-)TSST, as well as across the sitting and standing conditions

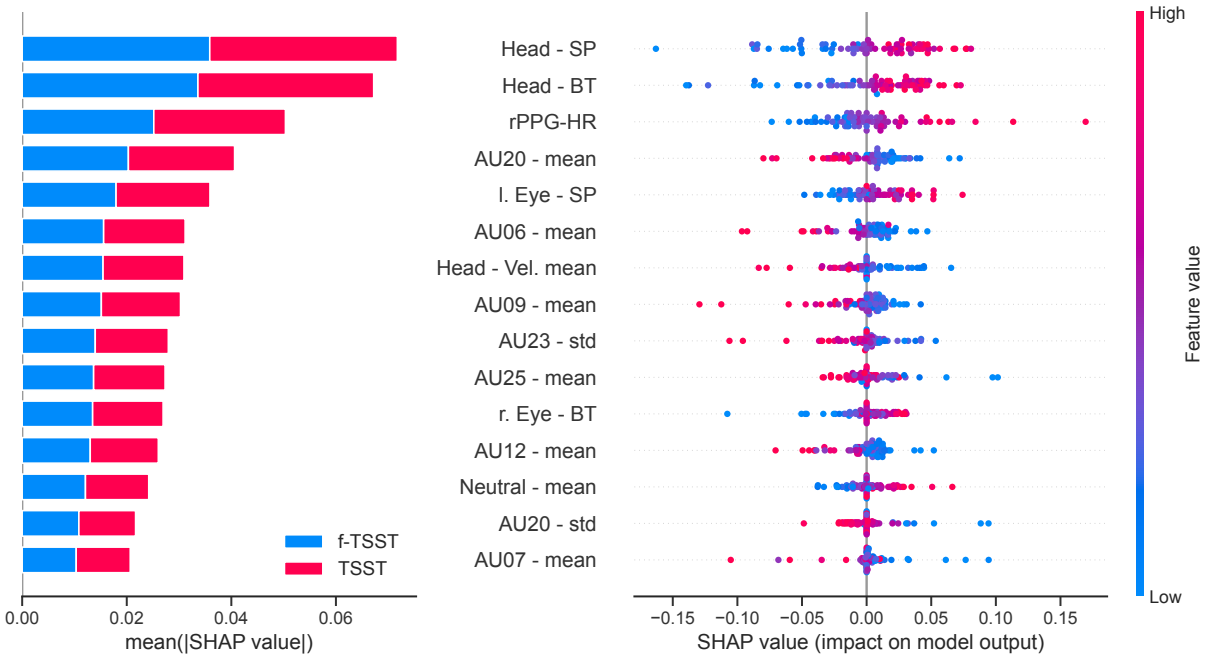

Figure 4.24: SHAP values of the multimodal approach for *Math*: Feature importances were determined using SHAP values calculated across the cross-validation folds of model evaluation. A higher absolute SHAP value signifies greater influence on the model's classification output. Positive SHAP values correlate with an increased likelihood of predicting the positive class, namely TSST. Note: SP = Static Periods, BT = Below Threshold, r.=right, l.=left.

Chapter 5

Discussion

This thesis was designed to investigate the feasibility of detecting acute psychosocial stress through video recordings by analyzing facial expressions, body movements, and gaze behavior, and to examine their correlations with traditional stress markers. Additionally, it aimed to validate advanced rPPG models under real-world conditions characterized by variable heart rates, speech activity, and head movements. The study also assessed the impact of rPPG-HR on the accuracy of stress predictions to better understand the potential and limitations of using non-invasive methods for stress detection in real-world scenarios.

The EmpkinS-TSST study aimed to induce acute psychosocial stress through the TSST while offering a stress-free control condition in a comparable scenario, the f-TSST. Results indicate that the TSST successfully induced acute psychosocial stress which is supported by biological markers such as a stronger HPA axis activation in form of higher cortisol levels, as well as a higher SNS activation indicated by higher HR levels during the TSST. Further, the responses of perceived negative affect and state anxiety were higher in the stressful condition. All responses to acute psychosocial stress are in line with existing literature [Goo17] [Lab19] [All17]. Conversely, the f-TSST did not activate the HPA axis or the SNS, and both perceived negative affect and anxiety were lower, consistent with findings from previous studies [Wie13].

The findings of this thesis prove three of four formulated hypotheses, demonstrating the efficacy of video-based digital biomarkers in distinguishing stress-related states. Specifically, Hypothesis 1 was verified, revealing that digital biomarkers can differentiate between stressed TSST and non-stressed f-TSST states with a classification accuracy of 73.3±5.5%. Critical to this finding was the role of facial expressions, body movements, and gaze behavior, which were all identified

by explainable AI algorithms as key markers influencing the classification output across different phases of the stress test.

For Hypothesis 2, SBMLR using digital biomarkers achieved an R^2 of 64%. However, the incorporation of PCA components in SBMLR added complexities that affected its generalization and interpretability. In contrast, machine learning-based regression yielded a minimal R^2 of 0.02 ± 0.15 for the cortisol slope S1 to S4. Given the limited success of the machine learning approach, this hypothesis was rejected due to the inability to effectively predict continuous values of HR, cortisol, and subjective stress scores from digital biomarkers alone.

Hypothesis 3 was validated under the premise that both conventional and DL-based rPPG models would exhibit decreased performance in more dynamic, naturalistic environments compared to controlled settings. This was evident as the models tested on the f-TSST dataset underperformed, especially in conditions of high movement, variable heart rates, and during speech tasks, aligning with the hypothesized challenges in naturalistic data application.

Finally, Hypothesis 4 was supported, illustrating that the integration of rPPG-derived heart rate measurements with behavioral digital biomarkers notably enhances the robustness of stress prediction models, particularly noted during the math phase of testing with an improved accuracy of 77.3 ± 6.5%. This enhancement underscores the potential of combining physiological and behavioral data for more robust stress detection models.

The thesis demonstrates the potential and challenges of using video-based digital biomarkers and rPPG models for stress detection, highlighting the critical need for rigorous methodological practices to ensure enhanced reliability and generalizability. Detailed examinations of the results for each hypothesis are provided in the following subsections, which will further clarify their broader scientific implications.

5.1 Aim 1: Using Digital Biomarkers to Predict Stress States

When observing the influence of acute stress on each digital biomarker group individually, results suggest that the stress response could be quantified by more neutral facial expressions, less upper body movement, and a more static gaze behavior throughout the TSST.
In the f-TSST, increased happy emotions and less neutral expressions were observed, accompanied by an overall higher AU intensity compared to the standard TSST. Specifically, AU06 and AU12, linked to happy expressions, were more active, aligning with findings that higher activity in AU12 correlates with lower cortisol increases [Bla23]. Additionally, AUs like AU09 (Nose Wrinkler),

AU14 (Dimpler), AU20 (Lip Stretcher), and AU25 (Lips Part) showed greater activity in the f-TSST. While these could be associated with emotions like fear, disgust, and contempt in conjunction with other AUs, it is important to note that these measurements, influenced by speaking behavior and determined frame-by-frame, lack temporal context and may primarily indicate increased speaking during the f-TSST [Sha13]. Without analysis of the speech data, further validation of these findings through estimated speech lengths or speech rates remains unconfirmed.

Significant differences in static body movements were noted, with fewer head movements observed in the TSST compared to the f-TSST, reflecting Richer et al.'s findings of reduced upper body movements using inertial sensor-based motion capturing suits [Ric24a]. These results highlight the feasibility of using video recordings to assess acute stress impacts on body movement. The analysis of mean head and shoulder velocities showed slower movements in the TSST, though these differences were not statistically significant. Video-based movement quantification can be problematic due to the distance between the camera and the participant, which may lead to inaccurate estimations of movement, as noted by [Thi10]. Additionally, pose estimation from video faces challenges such as reliable detection in conditions of occlusion and depth ambiguity. This complicates the analysis further, as the absence of the third dimension adds complexity to the interpretation of spatial relationships [Zhe24]. Continued research is needed to develop more sophisticated, relative movement features that address these challenges in video-based assessments.

Observations regarding gaze behavior revealed that only expert features indicated significantly fewer eye movements during the TSST, suggesting a more static gaze behavior. This finding is consistent with Herten et al., who observed longer and more frequent fixations on central objects when using eye-tracking glasses in the TSST compared to the f-TSST [Her17]. However, the EmpkinS-TSST study, which relied solely on video recordings, did not include gaze calibration prior to the (f-)TSST. Such calibration would have been essential for analyzing gaze times on committee faces versus the surroundings to assess gaze avoidance, a behavior found to be increased in the TSST in a previous study [Her17] [Che20]. Additionally, no significant differences were observed in pupil diameter between the TSST and f-TSST. In contrast, Guy et al. reported an increase in pupil diameter following the TSST when using eye-tracking glasses [Guy23]. Considering that pupil diameter varies between 2-8 mm and is sensitive to illumination changes, these subtle changes are challenging to detect solely with video cameras [Wan18].

The findings support previous research that has individually examined the relationship between acute stress and specific groups of digital biomarkers, such as facial expressions, body movement, and gaze behavior, confirming that each category is independently linked to stress. These results underscore a significant interplay between behavioral, endocrinological, and motoric systems during acute stress events. This interaction supports the integration of behavioral information as additional biomarker alongside established psychobiological markers, offering a more comprehensive view of the human stress response.

The sole use of video-based digital biomarker to detect acute psychosocial stress using machine learning supports this conjecture. A maximum accuracy of 73.4±7.7% was achieved to detect whether an individual was exposed to acute stress or not. This accuracy was lower than the ones observed in previous studies. For instance, Aigrain et al. attained a F1 score of 85%. However, they not only used video-based features but also integrated physiological markers such as blood volume pulse and HR measures into their model, as well as lacked a control condition [Aig18]. Zhang et al. achieved a detection accuracy of 85.42% in a deep learning based approach where participants had to answer questions or watched a neutral video sequence [Zha14]. Viegas et al. reported an accuracy of 81.1% in classifying stress from neutral tasks solely using AU [Vie18]. In contrast to the above named approaches, in this thesis nested cross-validation was employed for hyperparameter tuning and model evaluation, aiming to provide a less biased estimate of the classification pipelines' generalization performance [Vab19].

Feature importances from the machine learning model revealed that the top ten features were a mixture of head movements, facial expressivity, and gaze features. This suggests the model effectively differentiates between the TSST and f-TSST by recognizing behavioral differences using different digital biomarker groups. The machine learning approach offers an advantage over inferential analysis by using a combination of biomarkers, rather than relying solely on one type of behavioral feature. This integration helps to address challenges such as the impact of speech on facial expression accuracy and changes in participant-camera distance affecting movement and gaze measurements. Thus, the model effectivly handles the complex behavioral changes induced by acute psychosocial stress, illustrating the benefits of leveraging multiple video-derived features.

A deeper analysis of the phases revealed that the number of significant features remained consistent overall. However, only facial expression features exhibited a larger effect size during the *Talk* phase, while during the *Math* phase facial expression, upper body movements, and gaze

features showed larger effect sizes. The classification results mirrored this distinction: despite similar accuracies across phases, the feature importances diverged notably.

Moreover, in the *Talk* phase, the three most significant features were all related to the rate of change in the expression of happiness. This suggests that both the amount and the variations of smiling — how frequently it changes — were key indicators of a non-stressed state during speaking. Notably, other action units not associated with positive emotions were also more active in the non-stressful condition. This reliance on facial expression features for classification likely arises from two factors. First, action units may be influenced by speaking behavior, with facial muscles engaging more actively during speech [Sha13]. Alternatively, a possible hypothesis is that participants' facial muscles were more relaxed and expressive during the non-stressful *Talk* phase of the f-TSST and become tenser during the stressful TSST, due to variations in stress levels. Further research is necessary to assess the reliability of current methods used to detect facial expressivity and action units, especially when influenced by speaking.

Contrarily, analyzing solely the *Math* phase revealed that the most important features consist of a combination of movement, gaze, and facial expression features, with the top two being upper body ones. Unlike the *Talk* phase, the most influential facial expression features during the *Math* phase did not relate to positive emotions. This discrepancy suggests that these features might have been influenced by speaking behavior, similar to observations from the *Talk* phase. As a result, they may reflect the extent of speaking or the relaxation of facial muscles during speech rather than emotional states. This indicates that social-evaluative stress may impact behavior differently across the two phases, highlighting a key area for future research to explore the effects of acute psychosocial stress on facial expressions, body movements, and gaze behavior in varied scenarios.

5.2 Aim 2: Using Digital Biomarkers to Predict Continuous Stress Measures

When examining the relationship between changes in facial expressions, speech, movement, and traditional biological and psychological stress markers, the highest explained variance of 64% was observed in predicting the stress response slope from baseline to peak following the (f-)TSST. Despite these promising results, the model, derived from SBMLR, may not generalize well due to lack of validation on a separate test set, and the use of PCA complicates result interpretation since PCA components do not directly correspond to specific observed behavioral changes. In related research, Lasselin et al. achieved a comparable explained variance of 60% using SBMLR

to predict motion alterations based on variables like body temperature, sickness symptoms, back pain, and IL-6 concentrations [Las20]. This indicates a significant yet complex link between physiological changes and motion alterations that warrants further investigation.

Further, only cortisol achieved positive R^2 values in the ML-based regression models, underscoring the challenges in effectively predicting psychological and physiological responses using current machine learning techniques. The negative R^2 values observed for both Positive and Negative Affect Scores of the PANAS and mean HR highlight the limitations of the existing feature set in capturing the nuanced behaviors associated with stress responses. To enhance model performance, future work should consider the integration of more expert features that better capture human behavioral dynamics. Additionally, incorporating voice prosody features may provide a notable boost to the prediction of stress-related biomarkers, potentially leading to more robust models capable of handling the complexities of real-world stress detection.

5.3 Aim 3: rPPG Validation under Real-World Conditions

To develop a more comprehensive understanding of acute stress responses, it is essential to consider internal physiological changes. rPPG offers a promising methodology by enabling non-invasive HR estimation from facial videos, thereby eliminating the need for physical contact [All07]. Another aim of this thesis was to validate current state-of-the-art rPPG methods, initially on benchmark datasets and subsequently on the more real-world TSST datasets UBFC-PHYS and EmpkinS-TSST. Both datasets include challenging factors such as varying heart rate levels, speech sequences, and head movements.

Regarding the first analysis on the benchmark datasets, the best performance across all rPPG models was observed on UBFC-rPPG, followed by PURE. This aligns with previous findings, as both datasets are highly controlled for movement, illumination, and heart rate, making them ideal for testing models under stable conditions without challenges such as motion, speech, or lighting changes [Ni21] [Xia24]. However, the performance metrics reported in this thesis are slightly lower than those in prior studies. Specifically, this study's findings regarding the PURE dataset validate those of Liu et al., who noted that conventional methods like LGI and POS outperformed the DL-based model TSCAN, which was trained on UBFC-rPPG [Liu23]. For UBFC-rPPG, the lowest MAE was achieved using POS, whereas Liu et al. reported the best performance with TSCAN. This discrepancy might be due to the different postprocessing approaches. This thesis employed a finer HR frequency resolution of 3 bpm, compared to the 12-8.5 bpm resolution used

by Liu et al., which could contribute to the slightly poorer performance observed here [Liu23]. Additionally, the benefit of using a higher HR resolution is underscored when models are evaluated on datasets like EmpkinS-TSST, which provides HR data derived from ECG. Comparisons made with predictions at different HR frequency resolutions often result in larger MAEs.

In the COHFACE dataset, no model achieved a MAE lower than ten bpm. In contrast, the study by Liu et al. once again demonstrated better performance [Liu23]. This discrepancy might be as well attributed to the finer heart rate resolution used in the thesis. Additionally, while the PURE and UBFC-rPPG datasets feature similar heart rate ranges, COHFACE encompasses a broader spectrum. Given that the deep-learning models were primarily trained on the PURE or UBFC-rPPG datasets, their tendency to overestimate HR in COHFACE is not surprising. This illustrates the significant impact that the choice of training dataset can have on the predicted HR range [Xia24]. Moreover, conventional methods also showed limited success on COHFACE, potentially due to video quality degradation from heavy MPEG-4 Visual compression, which may corrupt the rPPG signal [McD17].

When evaluating rPPG models using the TSST datasets, the models demonstrated markedly better performance on the UBFC-PHYS dataset compared to the EmpkinS-TSST dataset. Notably, for UBFC-PHYS, the performance of deep-learning models remained consistent regardless of whether the models were trained on PURE or UBFC-rPPG which is in line with previous findings [Sab23]. Conversely, for the EmpkinS-TSST dataset, models trained on UBFC-rPPG outperformed those trained on other datasets.

One potential explanation for the observed performance disparities could be attributed to the design of the UBFC-PHYS dataset, which was specifically developed for rPPG benchmarking and emotion recognition. This dataset benefits from superior video quality and enhanced lighting conditions, factors that could contribute to the differences in model performance [Sab23].

Further analysis across the specific phases of *Pause*, *Talk*, and *Math* revealed performance variations. For both TSST datasets, the lowest MAE was achieved during the *Pause* phase, followed by the *Math* phase, with the *Talk* phase showing the worst performance. Previous studies have indicated that rPPG models struggle with factors like head movement, changes in illumination, and varying heart rate levels [Yan22] [Di 24]. The *Pause* phase, which requires participants to remain calm and face the camera, naturally presented the least challenging conditions for rPPG models, as evidenced by the lowest MAE observed. Comparatively, the *Math* phase involved less movement than the *Talk* phase, where higher facial activity likely due to more frequent speaking was noted. The increased movement and speech during the *Talk* phase made it the most chal-

lenging, as confirmed by the lower performance metrics observed for both the UBFC-PHYS and EmpkinS-TSST datasets.

When comparing the predicted heart rate levels between TSST and f-TSST, it is interesting that the model's performance is lower in the f-TSST than in the TSST. In the inferential digital biomarker analysis more upper body movement and a higher facial expressivity was observed in the f-TSST. Since more movement and more speech parts pose more challenges for the model, it could be argued that these performance disparities were unexpected. However, one important factor which is often overlooked in the current rPPG-benchmark datasets are the HR ranges. This shows that higher as well as lower HR ranges pose another challenge for the rPPG models [Di 24].

In summary, rPPG models are able to accurately predict the HR under highly standardized lab settings. However, the performance decreases when rPPG models face certain challenges, such as upper body movement, illumination changes, various heart rate levels, speaking, or different skin colors [Di 24]. One reason for that is the lack of benchmark datasets which incorporate these challenges. Although numerous datasets exist for the evaluation of rPPG methods, the availability of high-quality, open datasets remains constrained. Presently, the most widely utilized datasets among researchers are UBFC-rPPG, PURE, and COHFACE, which predominantly address motion artifacts and illumination changes [Xia24]. Nonetheless, these datasets do not account for several specific factors, including variations in heart rate levels, speaking parts, and environmental conditions. This omission poses significant challenges in assessing the effectiveness of various rPPG methods. Therefore, more open-source datasets are needed for training and evaluation which include the challenges and thus make allow rPPG models to tackle them.

The challenges associated with open resources extend beyond benchmark datasets to include accessibility issues with code for various learning-based methods. The availability of open-source code is crucial for both established researchers and newcomers in the field. Often, the absence of accessible code or clear reproducibility guidelines limits the ability to evaluate and compare methodological performance effectively [Xia24]. Furthermore, there is an ongoing need for enhancements and updates to open-source toolboxes that support efficient training and testing of models. Although the rPPG-toolbox strives to stay up to date, it currently offers only five pre-trained models on the PURE and UBFC-rPPG datasets [Liu23].

Additionally, researchers frequently do not release optimal sets of hyperparameters or model weights, complicating efforts to replicate results. The performance of deep learning models is also

non-deterministic, influenced by factors such as the type of graphics card used and the version of the CUDA toolbox, details that are often omitted in publications or repository documentation [Pha20]. Improvement in this area requires a collective effort from the research community to establish open-source benchmarking platforms. Such platforms would facilitate the uploading of pre-trained models and enable benchmarking across various datasets that present diverse challenges. This approach would significantly enhance transparency and reproducibility in the field [Xia24].

Overall, deep learning based models do not fully outperform conventional methods. They also have great advantages since they do not rely on training. Further, in this study conventional methods achieved comparable results while not relying on training data since they are unsupervised models. Consequently, unsupervised deep learning methods should be further investigated, as they can overcome the reliance on real labels in supervised methods and facilitate practical applications.

Part of the thesis focused on validating state-of-the-art rPPG models and identifying challenges within the more naturalistic EmpkinS-TSST dataset. Despite substantial potential for enhancing rPPG model performance, inferential analysis indicated that the predicted HR was significantly higher in the *Math* and *Pause* phase of the TSST conditions compared to f-TSST. This resembles the results of the recorded ground truth and as mentioned earlier, indicates the activation of the sympathetic nervous system which provides valuable insights into individual stress levels.

However in the *Talk* phase, the actual HR measured was significantly higher in the TSST compared to the f-TSST, yet this distinction was not mirrored in the predicted HR by the rPPG model. This discrepancy led to subsequent analysis to determine whether using a single physiological proxy could improve the prediction accuracy of psychosocial stress responses. As a result, the HR predictions from TSCAN, the highest performing model trained on the UBFC-rPPG dataset, were integrated as an additional digital biomarker. This integration aimed to enhance the detection of stress states through a multimodal approach, leveraging digital biomarkers to provide a more comprehensive assessment of stress responses.

5.4 Aim 4: Using a Multimodal Approach to predict Stress States

Integrating remote photoplethysmography-derived heart rate (rHR) with behavioral digital biomarkers is hypothesized to enhance the predictive accuracy and robustness of stress detection models. Using digital biomarkers across the (f-)TSST, the overall performance remained roughly the same

but the standard deviation across the different folds was lower with an accuracy of 73.3±5.6%. This shows that adding the rPPG-HR to the model improves the robustness across the CV folds. This was also confirmed by a deeper investigation of feature importances which revealed that the rPPG-HR was among the top 10 features to have an impact on the model output. The rest of the pattern stayed the same as before.

Regarding the *Math* phase, best overall performance was observed with an accuracy of 77.3±6.5% and an F1 score of 79.3±5.1%. When analyzing the feature importances, rPPG-HR was the third most important feature. This shows how a single HR proxy can already boost the performance of the classifier and allow a more accurate prediction of the stress states.

Nothing changed for the stress prediction of the *Talk* phase. Therefore, rPPG-HR was not among the top features which was also expected since by simply comparing the predicted HR during *Talk* across the f-TSST and TSST, no difference was found. This shows the disparity between the different conditions and that the additional HR did not help to boost the performance since the rPPG model was not able to predict higher HR levels during the *Talk* phase accurately. The *Talk* phase was also the most challenging condition for the model, therefore the results align with the expectations of current rPPG-HR predictions. It shows once again that it is currently still hard to correctly classify higher heart rates when they are paired with other challenges such as movement or talking [Di 24].

These results contrast with previous studies employing rPPG-HR. Morales-Fajardo et al. assessed academic anxiety in fifty-six undergraduates using video recordings and STAI scores during regular and exam *Math* classes, relying on demographic data to classify anxiety [Mor22]. Their study showed an accuracy increase from 86% to 96% when incorporating rPPG-HR data, though it lacked validation for the HR results. Similarly, in stress state detection on the UBFC-Phys dataset, an accuracy of 85.45% was achieved using only rPPG-HR features, with higher accuracy noted when compared to ground truth PPG or EDA signals. However, their claim of analyzing stress via a modified version of the TSST raises concerns as they did not explicitly report results for their stress measure PPG and EDA, potentially confusing different mental loads with actual stress states [Sab23]. McDuff et al. reported an 85% accuracy in distinguishing rest from cognitive stress states using only rPPG-derived HR, HRV, and breathing rates. However, their study was limited by a small sample size of 10 participants and lack of cross-validation, suggesting potential overfitting [McD14].

5.5 General Discussion and Limitations

The challenge of defining "stress" precisely in research persists, notably due to the ambiguity surrounding its characterization [Epe18]. This thesis focuses on acute psychosocial stress, known for significantly activating the HPA axis [Dic04]. This contrasts with studies that depend on unobtrusive measures, which often fail to measure HPA axis activity and rely heavily on subjective stress assessments. Unlike other studies that induce cognitive load, this research uses protocols like the TSST to elicit robust physiological stress responses [Kir93].

The use of the f-TSST, which does not activate the HPA axis, underscores the complexity of defining and measuring stress with a single method or biological marker. Often, studies use a preparation phase to compare stressful and non-stressful conditions, introducing variability as different types of behaviors are measured [Aig18] [Lup14]. Richer et al. highlighted the ambiguity of stress labels in psychosocial stress research, emphasizing the need for further validation [Ric24a]. Norden et al. investigated classification results using various stress labels. However, their study was limited as it only included male participants [Nor22b]. Future work is necessary to clarify disparities in stress labels to better understand the validity of digital biomarkers in assessing stress states.

Moreover, the order in which the conditions were administered, either the TSST or f-TSST first, may have influenced participant behavior due to the novelty of the scenario on the initial study day — a factor also highlighted by Richer et al. Higher effect sizes for both cortisol levels and self-reported measures were observed when the TSST was administered first, suggesting a more pronounced stress response due to the order of conditions [Ric24a]. Additionally, classification performance was better when the TSST was first, particularly for features calculated over both phases, although no differences were observed during the individual phases. Conversely, integrating rPPG-HR into the model resulted in higher accuracy when the f-TSST was administered first. This improvement could be attributed to the relative stability of rPPG-HR measurements, since the effect size of heart rate levels seemed unaffected by the condition order. Future research should continue to explore how the order of conditions impacts individual behaviors.

In the EmpkinS-TSST, participants were equally randomized between sitting and standing postures. Contrary to the effects observed with condition order, participants who were sitting had higher prediction confidence during the (f-)TSST and the *Talk* phase, although this trend did not extend to the *Math* phase. Interestingly, cortisol levels did not show variation in effect sizes based on posture. Incorporating rPPG-HR data into the prediction model significantly improved classification accuracy for sitting participants during the (f-)TSST and the *Talk* phase, indicating

that posture has a measurable impact on behavior and classification performance. In contrast, classification accuracy decreased for standing participants during the (f-)TSST, but increased during the *Math* phase. This discrepancy may be due to the rPPG-HR data being more accurately predicted for sitting participants, who likely moved less than those standing. Further studies with larger sample sizes are necessary to explore the influence of body posture on physiological responses and stress detection more comprehensively.

Regarding gender differences, the predictive models included more female participants (21 women vs. 16 men). Throughout all phases, female participants were more likely to be correctly classified than their male counterparts, except during the *Math* phase when the integrated resting heart rate (rPPG-HR) was considered. No significant differences in effect sizes for cortisol levels and perceived stress levels were observed between men and women. One possible explanation for the disparate classification results could be that the imbalanced gender distribution in the sample influenced the outcomes. This could potentially support the hypothesis that women and men exhibit different behavioral stress responses, as behavioral markers were primarily used for classification. Adding rPPG-HR to the model resulted in higher classification accuracy for male participants, which could be due to more pronounced differences in heart rate levels between the f-TSST and the TSST. This is further supported by overall higher effect sizes of heart rate measures for men. However, due to the small and imbalanced sample size, future research should further investigate these gender-related behavioral differences in response to stress.

Together, the impacts of condition order, participant posture and gender on response measures demonstrate the complex and individualized nature of physiological and behavioral responses to psychosocial stress. This complexity is mirrored in the SHAP values, which vary among participants. A detailed analysis indicated that facial features influenced the outcomes during the *Talk* phase, whereas a broader range of features, particularly head movements, played a critical role during the *Math* phase. These insights support the need for further development of each digital biomarker group to more precisely reflect the stress response to psychosocial stress.

This exploratory study offers an initial look at the use of digital biomarkers and the persistent challenges in the field. Segmenting the analysis by condition order and posture (sitting vs. standing) has reduced the study's power due to a smaller sample size. Although this approach has uncovered interesting behavioral patterns between the phases *Math* and *Talk*, it underscores the necessity for larger samples to firmly establish conclusions.

Particularly concerning facial features, the study revealed complexities in measuring facial intensity, which could be attributed either to speaking or inherent facial expressions. The AUs

themselves are prone to errors as facial expression recognition is evaluated frame-wise and not adjusted for speaking parts, pointing to a need for newer, temporal algorithms that can more accurately differentiate these factors [Sha13]. Current state-of-the-art models, such as those provided by PyFeat [Che21], demonstrate varying levels of detection accuracy. For instance, F1-scores for happiness (0.77), AU06 (0.71), AU12 (0.78), and AU25 (0.84) are relatively high, indicating reliable detection of these expressions. However, F1-scores for other AUs and emotional values range from 0.25 to 0.64, suggesting that while the increased positive emotions observed in the f-TSST are reliable, individual action units might not be as dependable. This emphasizes the importance of considering emotion levels and AUs in conjunction with other digital biomarkers.

Moreover, there is considerable potential in further developing expert features for gaze behavior and upper body movements, which have been identified as highly ranked features. Enhancing the characterization of upper body movements could be achieved by integrating more IMU-based expert features, such as frequency-based features by Richer et. al [Ric24a], and refining the calculation methods for video-based movement features. For pose estimation, open-source platforms such as AlphaPose [Fan23] or OpenPose [Cao21] could also be considered. Additionally, incorporating more eye-tracking features through a brief calibration before tests could provide valuable insights into gaze behavior, such as gaze avoidance, which previous studies have shown to offer substantial insights into participants' mental states [Che20] [Vat21]. This exploration and detailed feature extraction could leverage the potential of stress measurements.

Another limitation of this study is the inability to utilize voice features due to poor audio quality. This omission represents a significant gap, as previous research has demonstrated the importance of vocal indicators in predicting acute stress [Oes23]. Incorporating voice features could have provided a more comprehensive analysis, enabling an examination of the interplay between facial expressions, speech patterns, body movements, and established stress markers like cortisol levels and subjective stress perceptions. Although this study did not originally aim to use video recordings to predict psychosocial stress, the results remain promising given that the differentiation between stress states was achieved solely through video-based features. Future research should aim to address this gap by integrating voice analysis to enrich the understanding of stress markers.

A further constraint of the current analysis is the assessment of facial features across entire phases rather than temporally, potentially overlooking subtle changes. Although the study attempted to model temporal changes in movement and gaze by identifying static behaviors, there is still significant room for improvement. Developing more advanced temporal digital biomarkers

could more effectively capture initial reactions to acute stress, especially in situations that evoke strong emotions where immediate, uncontrolled behavioral responses are crucial before they can be masked or altered.

There is a compelling case for deploying digital biomarkers in less controlled, real-world settings to validate their effectiveness under challenges like camera movement, varying lighting, background noise, and the presence of multiple people. For instance, video recordings have already been effectively used in emergency departments to predict PTSD and depression symptoms in trauma survivors [Sch24]. In another pilot study, the severity of major depressive disorders and the response to antidepressant treatments were remotely measured using digital biomarkers. These biomarkers, including facial expressions, head movements, and speech patterns, were extracted from smartphone recordings [Abb21]. Similarly, a subsequent exploratory study utilized digital biomarkers, including HRV measures derived from rPPG, eye-tracking, and voice markers, to screen for mental health issues in children, predicting levels of depression, anxiety, and stress [Cho24]. These studies illustrate that digital biomarkers have the potential to function outside traditional lab settings, providing objective, unobtrusive, and scalable assessments. They could be particularly beneficial in mental health for the early detection, ongoing assessment, and monitoring of treatment responses for psychological disorders.

Regarding rPPG measures, performance varied across datasets due to challenges related to movement, illumination, and heart rate levels. Although the EmpkinS-TSST was not specifically designed for rPPG extraction, incorporating it as an additional digital biomarker enhanced the model's performance and robustness. However, the effectiveness of rPPG also depended heavily on the phase of the TSST, with higher heart rates and speaking phases posing the greatest challenges for rPPG prediction in the EmpkinS-TSST. Further complicating robust rPPG analysis, current datasets often lack diversity in speaking sequences, higher heart rate levels, and different skin colors. Therefore, raw rPPG-HR levels should be interpreted with caution, particularly when the specific challenges are known. Future work should focus on further validating current state-of-the-art rPPG models, developing more challenging benchmark datasets that better reflect the complexity of real-world scenarios, and creating an open-source platform for comparing model performances and sharing parameters.

Despite these challenges, there is significant potential in expanding rPPG applications. It is also promising that rPPG methods have been used to measure a wide range of other physiological parameters, such as breathing rate, blood volume pressure, and heart rate variability. Expanding

the use of rPPG to include these measures could provide a more holistic insight into participants' mental and cognitive states.

The findings underscore the potential of leveraging facial expressivity, upper body movement, gaze features, and rPPG-HR to assess acute stress reactions. This approach could more effectively differentiate between cognitive load and acute stress situations that involve social-evaluative stress. While extracting meaningful behavioral data during acute stress is challenging, this study represents a step toward a more comprehensive understanding of the human stress response. By establishing these markers as supplements to traditional psychobiological markers, additional insights into the bodily responses to acute psychosocial stress are revealed. Furthermore, since these markers can be more readily measured in naturalistic settings compared to traditional biomarkers, incorporating video-based digital biomarkers when assessing stress responses could enhance the understanding of the interplay between behavioral, physiological, and motoric processes during stress.

Chapter 6

Conclusion

This thesis explored the use of video-based digital biomarkers to detect acute psychosocial stress, using facial expressions, body movements, and gaze behavior. Additionally, it incorporated the feasibility of rPPG models under challenging conditions such as variable HR, head movements, and speech sequences. This work aimed to advance non-invasive, scalable acute stress detection technologies, enhancing the general understanding of the interplay between behavioral and physiological responses to stress.

The first aim of the thesis focused on analyzing the impact of acute stress on facial expressions, body movements, and gaze behavior. Findings indicated that acute stress led to more neutral facial expressions, reduced upper body movement, and more static gaze behavior. Using digital biomarkers throughout the (f-)TSST, the classifier detected acute stress exposure with an accuracy of 73.4±7.7%. A deeper analysis of the phases revealed that facial emotion features were the most influential ones for decision-making in the *Talk* phase, while in the *Math* phase, all feature groups (facial expressions, upper body movements, gaze behavior) contributed, with upper body movements being the most important ones. This finding suggests that different forms of acute stress differently influence human behavior, highlighting the need for further research to better understand behavioral stress responses.

The thesis identified a relationship between behavioral changes and traditional stress markers, noting connections between altered behavioral expressions and both biological and psychological markers. However, specific measurements such as cortisol levels, perceived stress, or HR levels could not be directly predicted using digital biomarkers. Incorporating voice prosody into the predictive model may enhance its feasibility. Further research is essential to better understand the links between behavioral adaptations, physiological changes, and perceived stress levels.

Another aim involved validating advanced rPPG methods in more real-world settings characterized by variable HRs, speech, and head movements. The findings demonstrated that while rPPG models perform well in controlled lab settings, their accuracy decreases under less constraint scenarios such as the EmpkinS-TSST due to increased movement, higher HR levels, and speaking parts. These challenges are amplified by the lack of diversity in current datasets regarding speaking sequences, HR ranges, and skin color differences, which can remarkably affect rPPG signal quality and interpretation. Despite these challenges, rPPG still offers a promising contactless tool for HR prediction, emphasizing the need for more robust models and comprehensive benchmark datasets that mirror the complexities of real-world environments. This underscores the necessity of creating an open-source platform to share parameters and compare model performances, fostering improvements in rPPG technology, and expanding its application to broader physiological assessments.

Finally, the thesis assessed how rPPG-derived HR influences the prediction of stress states using video-based digital biomarkers. Adding rPPG-HR to the predictive models generally enhanced both robustness and accuracy, particularly in the *Math* phase of the TSST, achieving an accuracy of 77.3% $\pm$6.5%. This suggests that rPPG-HR is a valuable addition to the set of digital biomarkers. However, performance varied across phases, with the rPPG models struggling the most during the *Talk* phase, which typically exhibits the highest HR levels and more speaking parts compared to the *Math* phase. These findings indicate that while rPPG models show promise, they still require further refinement to ensure reliability in real-world settings, and should be used with caution.

In summary, this thesis highlights the potential of using video-based digital biomarkers for acute psychosocial stress detection, especially when integrated with rPPG-derived physiological data. While challenges persist in accurately capturing and interpreting these digital biomarkers under varied and real-world conditions, the research enhances the understanding of the interplay between behavioral and physiological responses to acute stress. Future work should focus on refining video-based methods and expanding the scope of data collection to improve the robustness and practical applicability of stress detection technologies in real-world settings. The promise of video-based digital biomarkers lies in their potential to simplify the understanding and management of stress across various applications, including clinical, workplace, and everyday environments.

Chapter 7

Future Work

This thesis has demonstrated the potential of quantifying human behavior in the context of acute psychosocial stress using only video recordings. Future research should focus on refining the analysis of human behavior under stress by developing more sophisticated features for facial expressions, upper body movements, and gaze behavior. These advanced features will allow for a deeper understanding of the interactions between internal psychological states and physiological body signals, enhancing the granularity with which stress is measured and interpreted.

Facial emotion recognition, particularly during speaking tasks, deserves specific attention to determine whether changes in facial expressions occur due to the amount of speaking or to actual emotional shifts. Additionally, incorporating quick eye-tracking calibration in future studies will enhance the reliability of gaze behavior data, providing more accurate insights into where individuals direct their attention during stress. This precision in measuring gaze could illuminate new aspects of how stress affects attentional focus.

The proof of concept presented in this thesis has highlighted behavioral variations across different scenarios, such as the *Talk* and *Math* phases. Expanding upon this, a more extensive exploration of the response to acute psychosocial stress in varied scenarios is necessary. This will enable a more nuanced understanding of how stress manifests across diverse settings and help to tailor interventions more effectively.

Another future goal is to integrate various behavioral features into a comprehensive, multimodal approach that captures the temporal dynamics of human behavior. Considering the complexity of human behavior as an interplay of facial expressions, body movements, gaze behavior, and voice prosody, each evolving over time, future research should explore these temporal interactions to detect initial responses to acute stressors — before coping mechanisms and behavioral masking obscure these reactions. Deep learning models, particularly the Temporal Fusion Trans-

former [Lim21], are well-suited for this task due to their capability in handling time-series data and integrating different types of data for a holistic understanding of stress responses.

In terms of remote assessment of physiological signals, further steps should include the validation of rPPG models in more naturalistic settings, involving challenges such as participants speaking, diverse skin tones, and fluctuating HR levels. Moreover, enhancing the extraction of HRV features from rPPG data and exploring other remotely assessable physiological signals like breathing rate and blood pressure will broaden the scope of detectable physiological responses to stress.

As remote assessments of acute stress become more prevalent, it is essential to validate these emerging methodologies against traditional stress markers to ensure their accuracy and reliability [Nor22a] [Ric24b]. The use of video recordings as proxies for stress responses, particularly when combined with remote physiological markers like HR measurements, holds remarkable interest for fields such as remote patient care and virtual doctor visits. In these settings, they can offer valuable insights into patients' physiological and mental states, thus enabling more patient-centered care.

Finally, exploring stress responses in more natural environments outside the laboratory setting could improve the general understanding of the transition from acute to chronic stress [Roh19]. With the widespread availability of cameras in smartphones and computers, video-based stress detection offers a cost-effective and easily implementable solution for monitoring stress in daily life and work environments. This approach not only allows for the early detection of situations where individuals are repeatedly exposed to acute stress but also enables targeted interventions. By leveraging video-based digital biomarkers, there is the potential to disrupt the cycle where acute stress evolves into chronic stress, offering a promising approach for proactive mental health management. This expanded use could ultimately lead to more robust, real-world applications that extend beyond the lab, supporting preventive health measures and enhancing overall well-being.

Appendix A

Additional Figures

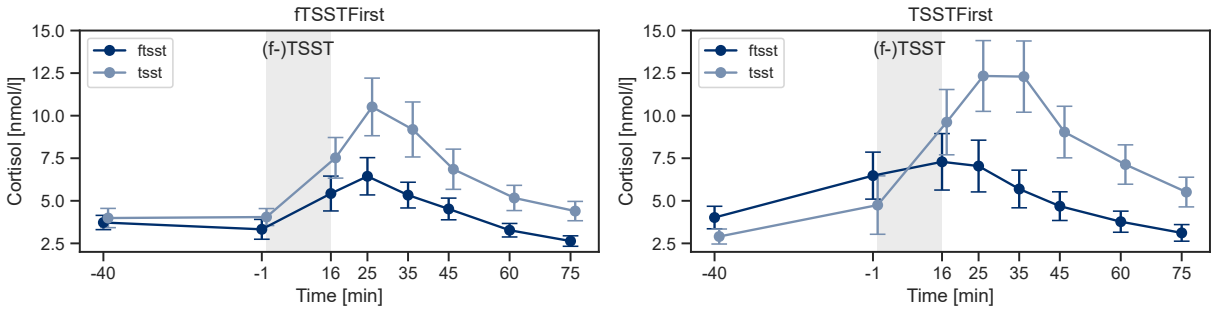

Figure A.1: Cortisol response per condition order; Mean ±SE over all participants

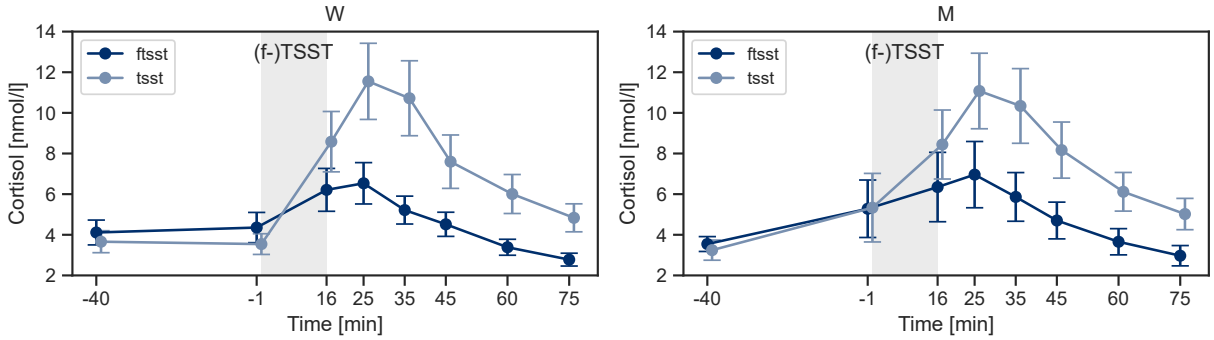

Figure A.2: Cortisol response per gender; Mean ±SE over all participants

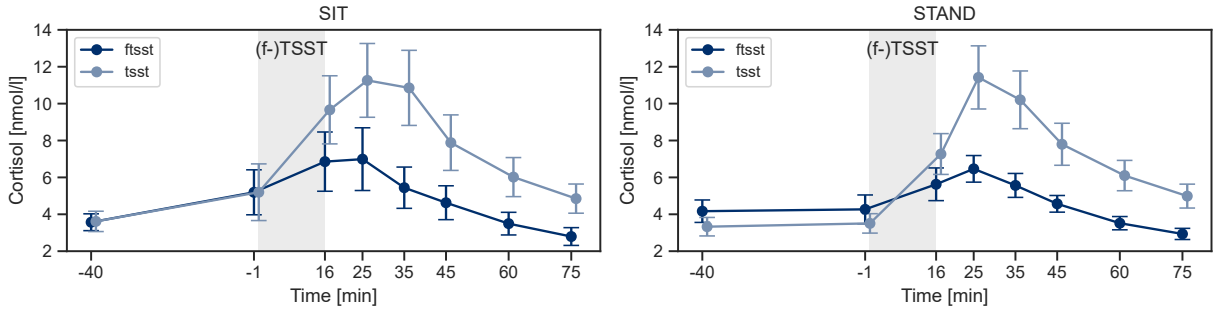

Figure A.3: Cortisol response per sitting and standing condition; Mean ±SE over all participants

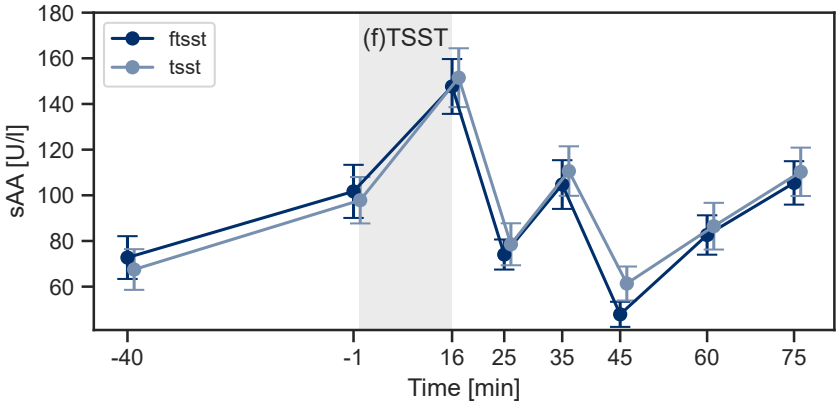

Figure A.4: sAA response during the (f-)TSST; Mean ±SE over all participants

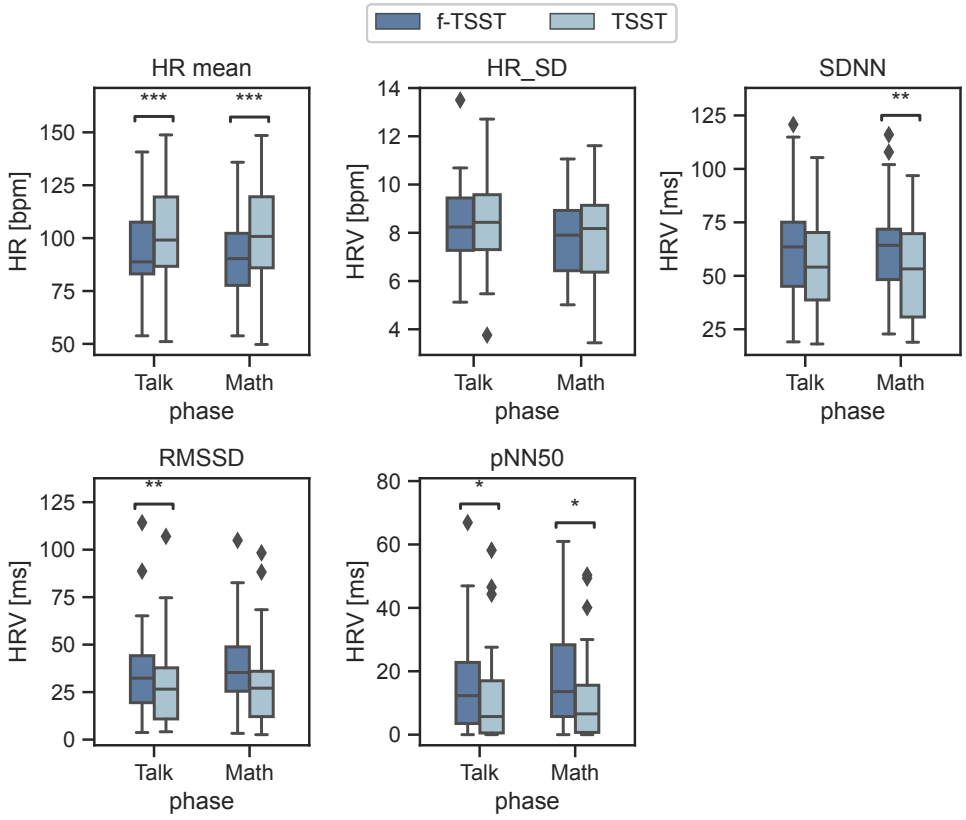

Figure A.5: HR and HRV results over all participants across the phases *Math* and *Talk*; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

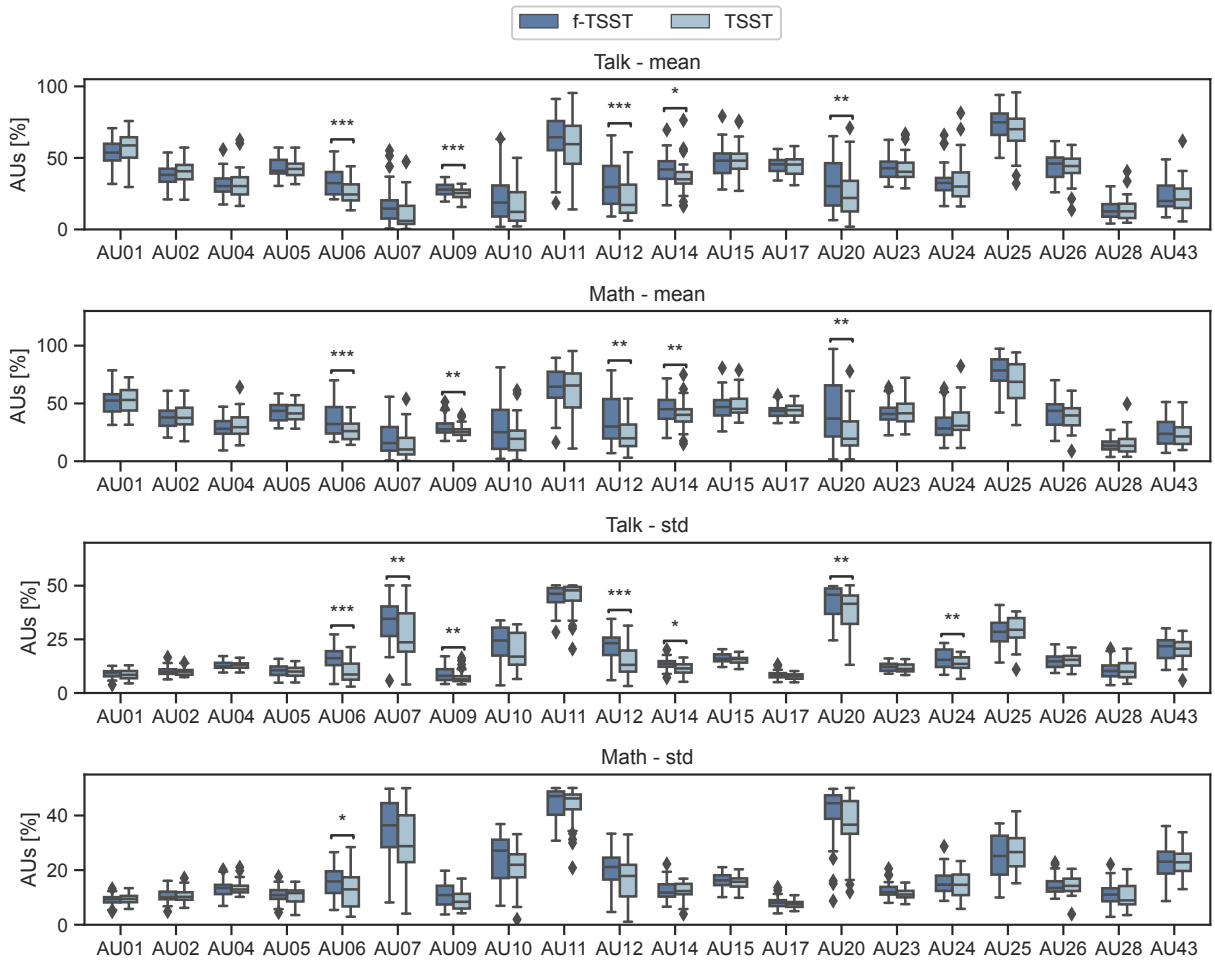

Figure A.6: Mean and standard deviation results for all AUs across all participants across the phases *Math* and *Talk*; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

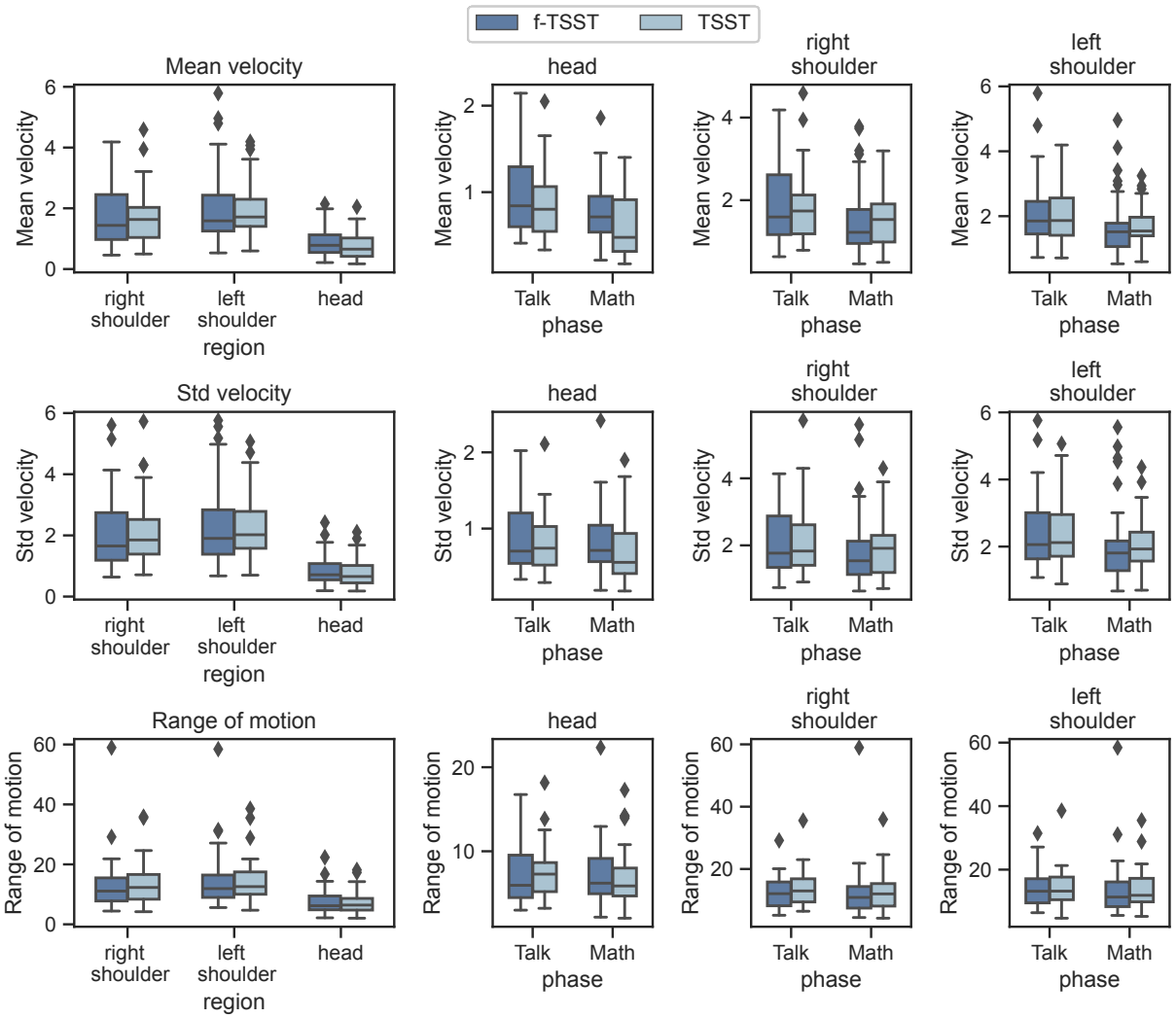

Figure A.7: Results of the generic head movement features for head, left and right shoulder during the (f-)TSST across all participants, as well as for the individual phases

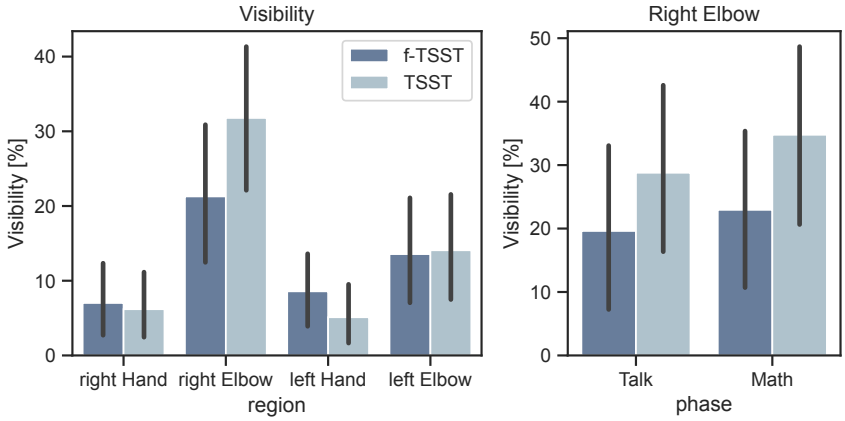

Figure A.8: Results of the visibility features for both elbows and hands during the (f-)TSST across all participants, as well as for the individual phases for the right elbow.

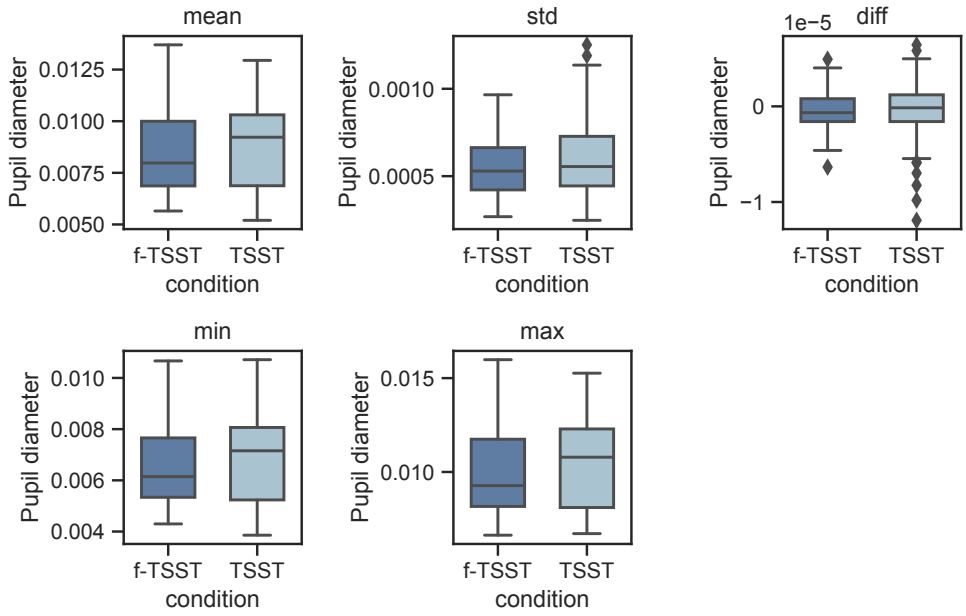

Figure A.9: Results of the pupil diameter features during the (f-)TSST across all participants. Note: the mean of left and right pupil is shown; diff: temporal difference

Appendix B

Additional Tables

Table B.1: t-test results of cortisol features for condition order; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Test Condition	Feature	t(43)	p	Hedges' g
TSST first	$aucG$	2.452	0.261	0.608
	$aucI$	4.308	0.005**	1.005
	$maxInc$	3.683	0.020*	1.010
	$maxVal$	2.370	0.307	0.545
	m_{S1S4}	4.389	0.005**	1.160
f-TSST first	$aucG$	1.889	0.734	0.512
	$aucI$	1.450	>0.999	0.401
	$maxInc$	1.721	>0.999	0.486
	$maxVal$	2.067	0.519	0.540
	m_{S1S4}	1.634	>0.999	0.485

Table B.2: t-test results of cortisol features for sit and standing conditions; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

test	Feature	t(43)	p	Hedges' g
SIT	AUC_G	2.070	0.516	0.478
	AUC_I	2.577	0.180	0.589
	Δc_{max}	2.439	0.242	0.637
	$maxVal$	1.909	0.707	0.407
	m_{S1S4}	2.848	0.099	0.754
STAND	AUC_G	2.447	0.263	0.727
	AUC_I	2.703	0.157	0.891
	Δc_{max}	2.722	0.151	0.861
	$maxVal$	2.702	0.157	0.832
	m_{S1S4}	2.572	0.205	0.860

Table B.3: t-test results of cortisol features for different gender identities; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

test	Feature	t(43)	p	Hedges' g
Men	AUC_G	1.705	>0.999	0.459
	AUC_I	2.177	0.458	0.573
	Δc_{max}	2.296	0.365	0.832
	$maxVal$	1.775	0.963	0.443
	m_{S1S4}	2.110	0.521	0.669
Women	AUC_G	2.603	0.166	0.617
	AUC_I	3.043	0.062	0.798
	Δc_{max}	2.806	0.106	0.678
	$maxVal$	2.597	0.168	0.602
	m_{S1S4}	3.243	0.039^{*}	0.889

Table B.4: t-test results of HR features for condition orders; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

test	Feature	t(41)	p	Hedges' g
TSST first	HR mean	3.681	0.015*	0.484
	HR SD	0.961	>0.999	0.177
	RMSSD	-3.464	0.024*	-0.450
	pNN50	-3.620	0.017*	-0.446
f-TSST first	HR mean	4.680	0.002**	0.502
	HR SD	-1.479	>0.999	-0.193
	RMSSD	-3.323	0.036*	-0.494
	pNN50	-3.144	0.053	-0.522

Table B.5: t-test results of HR features for sit and standing conditions; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

test	Feature	t(41)	p	Hedges' g
SIT	HR mean	4.401	0.003**	0.418
	HR SD	-0.572	>0.999	-0.088
	RMSSD	-2.997	0.071	-0.386
	pNN50	-3.199	0.045*	-0.419
STAND	HR mean	3.978	0.008**	0.570
	HR SD	0.466	>0.999	0.075
	RMSSD	-3.687	0.016*	-0.607
	pNN50	-3.280	0.039*	-0.598

Table B.6: t-test results of HR features for different gender identities; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

test	Feature	t(41)	p	Hedges' g
Men	HR mean	3.862	0.014*	0.633
	HR SD	0.312	>0.999	0.049
	RMSSD	-4.045	0.009**	-0.628
	pNN50	-3.665	0.021*	-0.633
Women	HR mean	4.384	0.002	0.408
	HR SD	-0.503	>0.999	-0.077
	RMSSD	-2.945	0.073	-0.382
	pNN50	-2.965	0.069	-0.394

Table B.7: t-test results of facial expression features; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Feature	t(37)	p	Hedges' g
fer intensity	7.434	$<0.001^{***}$	0.694
AU06 std	6.842	$<0.001^{***}$	0.780
AU12 std	6.413	$<0.001^{***}$	0.801
AU06 mean	6.284	$<0.001^{***}$	0.859
AU12 mean	5.715	$<0.001^{***}$	0.786
happiness mean	5.653	$<0.001^{***}$	0.716
happiness std	5.585	$<0.001^{***}$	0.733
AU20 mean	5.580	$<0.001^{***}$	0.717
AU09 mean	5.565	$<0.001^{***}$	0.766
AU14 mean	5.218	0.001^{**}	0.424
AU09 std	4.522	0.006^{**}	0.546
AU25 mean	4.491	0.007^{**}	0.596
AU07 std	4.389	0.009^{**}	0.574
neutral mean	−4.026	0.027^{*}	−0.537
AU10 std	3.896	0.039^{*}	0.397
AU24 std	3.744	0.060	0.436
AU07 mean	3.724	0.063	0.501
AU20 std	3.713	0.066	0.467
AU10 mean	3.562	0.101	0.403
AU15 std	3.400	0.158	0.439
AU17 std	3.324	0.194	0.390
AU23 std	3.221	0.258	0.478
AU14 std	2.683	>0.999	0.358
AU01 mean	−2.653	>0.999	−0.204
AU17 mean	−0.801	>0.999	−0.070
sadness std	−1.082	>0.999	−0.139
sadness mean	−0.955	>0.999	−0.144
neutral std	0.643	>0.999	0.118
AU01 std	0.175	>0.999	0.028
AU02 mean	−1.943	>0.999	−0.170
AU02 std	0.418	>0.999	0.050
AU04 mean	−1.021	>0.999	−0.151
fear std	−1.024	>0.999	−0.119
fear mean	−1.531	>0.999	−0.208

disgust std	0.320	>0.999	0.043
disgust mean	0.530	>0.999	0.099
anger std	−0.124	>0.999	−0.014
anger mean	−0.517	>0.999	−0.076
AU15 mean	−0.804	>0.999	−0.082
AU43 std	0.183	>0.999	0.016
AU28 std	0.622	>0.999	0.057
AU28 mean	−0.881	>0.999	−0.117
AU26 std	−0.309	>0.999	−0.031
AU26 mean	0.787	>0.999	0.093
AU25 std	−1.717	>0.999	−0.227
AU04 std	0.515	>0.999	0.074
AU05 mean	0.137	>0.999	0.010
AU24 mean	−1.586	>0.999	−0.130
AU05 std	1.362	>0.999	0.134
AU23 mean	0.361	>0.999	0.030
surprise mean	−1.898	>0.999	−0.234
AU11 mean	1.714	>0.999	0.216
AU11 std	0.004	>0.999	0.001
AU43 mean	1.516	>0.999	0.109
surprise std	−1.223	>0.999	−0.151

Table B.8: t-test results of upper body movement features; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Feature	**t(37)**	**p**	**Hedges' g**
head SP	−5.509	<0.001***	−0.728
head BT	−4.620	0.005**	−0.584
head Vel. mean	2.967	0.505	0.374
left Elbow Visibility	−0.109	>0.999	−0.017
left Hand Visibility	1.118	>0.999	0.198
left Shoulder BT	0.172	>0.999	0.027
left Shoulder Vel. mean	0.215	>0.999	0.040
left Shoulder SP	0.084	>0.999	0.014
left Shoulder Vel. std	0.185	>0.999	0.036
head Vel. std	1.816	>0.999	0.227
right Elbow Visibility	−1.275	>0.999	−0.254
right Hand Visibility	0.266	>0.999	0.046
right Shoulder BT	−1.686	>0.999	−0.282
right Shoulder Vel. mean	−0.162	>0.999	−0.032
right Shoulder SP	−1.332	>0.999	−0.218
right Shoulder Vel. std	−0.309	>0.999	−0.062

Table B.9: t-test results of gaze behavior features; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Feature	**t(37)**	**p**	**Hedges' g**
left Eye SP	-3.782	0.054	-0.522
right Eye SP	-3.279	0.220	-0.523
right Eye BT	-2.807	0.762	-0.417
left Eye BT	-2.520	>0.999	-0.328
left Pupil max	-2.190	>0.999	-0.312
right Pupil max	-2.122	>0.999	-0.304
right Pupil std	-1.662	>0.999	-0.266
right Pupil min	-2.010	>0.999	-0.265
right Pupil mean	-2.019	>0.999	-0.287
right Pupil diff	-0.351	>0.999	-0.065
right Pupil Blinks	0.427	>0.999	0.042
left Eye Vel. std	-0.157	>0.999	-0.022
left Eye Vel. mean	0.364	>0.999	0.046
left Pupil std	-1.467	>0.999	-0.221
left Pupil min	-1.627	>0.999	-0.222
left Pupil mean	-1.906	>0.999	-0.257
left Pupil diff	-0.144	>0.999	-0.026
left Pupil Blinks	1.571	>0.999	0.174
right Eye Vel. mean	0.148	>0.999	0.019
right Eye Vel. std	-0.668	>0.999	-0.093

Table B.10: t-test results of facial expression features for the phases Math and Talk; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Feature	Phase	t(37)	p	Hedges' g
AU06 std	Talk	7.534	$<0.001^{***}$	0.988
AU12 std	Talk	7.193	$<0.001^{***}$	0.984
fer intensity	Talk	6.105	$<0.001^{***}$	0.705
happiness mean	Talk	5.804	$<0.001^{***}$	0.742
AU06 mean	Talk	5.761	$<0.001^{***}$	0.884
happiness std	Talk	5.752	$<0.001^{***}$	0.859
AU12 mean	Talk	5.238	0.001^{**}	0.766
AU09 mean	Talk	5.165	0.002^{**}	0.676
fer intensity	Math	5.111	0.002^{**}	0.600
AU06 mean	Math	5.074	0.002^{**}	0.748
AU20 mean	Math	4.995	0.003^{**}	0.790
AU24 std	Talk	4.813	0.005^{**}	0.597
AU20 mean	Talk	4.747	0.006^{**}	0.524
AU09 mean	Math	4.707	0.007^{**}	0.680
AU12 mean	Math	4.683	0.007^{**}	0.718
AU14 mean	Math	4.638	0.009^{**}	0.386
AU09 std	Talk	4.440	0.016^{*}	0.545
AU07 std	Talk	4.434	0.016^{*}	0.614
AU20 std	Talk	4.366	0.019^{*}	0.495
AU14 mean	Talk	4.208	0.031^{*}	0.419
AU14 std	Talk	4.190	0.033^{*}	0.621
happiness mean	Math	4.158	0.036^{*}	0.619
AU06 std	Math	3.915	0.073	0.489
neutral mean	Math	−3.854	0.088	−0.607
AU09 std	Math	3.803	0.101	0.472
AU10 std	Talk	3.753	0.117	0.438
AU25 mean	Math	3.706	0.134	0.602
AU25 mean	Talk	3.661	0.152	0.488
AU15 std	Talk	3.623	0.169	0.573
AU12 std	Math	3.596	0.183	0.509
happiness std	Math	3.385	0.329	0.500
AU10 mean	Talk	3.375	0.338	0.378
AU17 std	Talk	3.322	0.391	0.439
AU23 std	Math	3.292	0.425	0.608
AU07 mean	Talk	3.272	0.448	0.460
AU07 mean	Math	3.199	0.547	0.476

AU07 std	Math	3.015	0.890	0.460
neutral mean	Talk	−2.997	0.935	−0.376
AU10 mean	Math	2.949	>0.999	0.382
AU01 mean	Talk	−2.929	>0.999	−0.267
AU02 mean	Talk	−2.811	>0.999	−0.293
AU43 std	Math	−0.355	>0.999	−0.035
AU43 mean	Talk	0.671	>0.999	0.060
AU43 mean	Math	1.566	>0.999	0.142
AU28 std	Talk	0.783	>0.999	0.073
anger mean	Math	−0.230	>0.999	−0.042
anger mean	Talk	−0.746	>0.999	−0.094
AU28 std	Math	0.318	>0.999	0.035
AU28 mean	Talk	−0.741	>0.999	−0.083
anger std	Math	−0.203	>0.999	−0.029
AU43 std	Talk	0.694	>0.999	0.069
AU01 mean	Math	−1.877	>0.999	−0.136
fear std	Math	0.345	>0.999	0.038
disgust mean	Math	1.332	>0.999	0.248
surprise mean	Talk	−1.892	>0.999	−0.236
surprise mean	Math	−1.393	>0.999	−0.188
sadness std	Talk	−0.687	>0.999	−0.098
sadness std	Math	−1.045	>0.999	−0.157
sadness mean	Talk	−1.003	>0.999	−0.161
sadness mean	Math	−0.648	>0.999	−0.111
neutral std	Talk	0.798	>0.999	0.164
neutral std	Math	0.356	>0.999	0.061
fear std	Talk	−2.092	>0.999	−0.284
AU28 mean	Math	−0.894	>0.999	−0.142
fear mean	Talk	−2.223	>0.999	−0.375
fear mean	Math	−0.733	>0.999	−0.084
disgust std	Talk	−0.232	>0.999	−0.031
disgust std	Math	0.698	>0.999	0.102
disgust mean	Talk	−0.671	>0.999	−0.122
anger std	Talk	−0.009	>0.999	−0.001
AU26 std	Talk	0.155	>0.999	0.015
AU20 std	Math	2.481	>0.999	0.372

AU26 mean	Talk	-0.062	>0.999	-0.008
AU11 mean	Math	1.393	>0.999	0.195
AU10 std	Math	2.512	>0.999	0.312
AU05 std	Talk	1.400	>0.999	0.160
AU05 std	Math	0.791	>0.999	0.096
AU05 mean	Talk	-0.524	>0.999	-0.052
AU05 mean	Math	0.791	>0.999	0.057
AU04 std	Talk	1.128	>0.999	0.157
AU04 std	Math	0.002	>0.999	0.000
AU04 mean	Talk	-0.573	>0.999	-0.074
AU04 mean	Math	-1.257	>0.999	-0.207
AU02 std	Talk	0.401	>0.999	0.053
AU02 std	Math	0.286	>0.999	0.036
AU02 mean	Math	-0.439	>0.999	-0.041
AU01 std	Talk	1.737	>0.999	0.247
AU01 std	Math	-1.341	>0.999	-0.240
AU11 mean	Talk	1.483	>0.999	0.210
AU11 std	Math	0.273	>0.999	0.048
AU11 std	Talk	-0.257	>0.999	-0.050
AU14 std	Math	0.470	>0.999	0.072
AU26 mean	Math	1.224	>0.999	0.169
AU25 std	Talk	-1.001	>0.999	-0.121
AU25 std	Math	-1.476	>0.999	-0.272
AU24 std	Math	1.831	>0.999	0.248
AU24 mean	Talk	-0.336	>0.999	-0.026
AU24 mean	Math	-1.925	>0.999	-0.216
AU23 std	Talk	2.018	>0.999	0.257
AU26 std	Math	-0.619	>0.999	-0.074
AU23 mean	Talk	1.435	>0.999	0.118
surprise std	Math	-0.904	>0.999	-0.122
AU17 std	Math	2.304	>0.999	0.300
AU17 mean	Talk	0.619	>0.999	0.055
AU17 mean	Math	-1.891	>0.999	-0.196
AU15 std	Math	2.435	>0.999	0.280
AU15 mean	Talk	-1.114	>0.999	-0.115
AU15 mean	Math	-0.311	>0.999	-0.038
AU23 mean	Math	-0.522	>0.999	-0.052
surprise std	Talk	-1.128	>0.999	-0.145

Table B.11: t-test results of upper body movement features for the phases Math and Talk; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Feature	Phase	t(37)	p	Hedges' g
head SP	Math	-5.455	0.001^{**}	-0.815
head BT	Math	-4.799	0.005^{**}	-0.690
head Vel. mean	Math	3.484	0.250	0.511
head SP	Talk	-2.560	>0.999	-0.423
head Vel. std	Math	1.493	>0.999	0.243
head Vel. std	Talk	1.033	>0.999	0.154
right Elbow Visibility	Math	-1.294	>0.999	-0.270
right Elbow Visibility	Talk	-1.181	>0.999	-0.225
left Elbow Visibility	Math	0.134	>0.999	0.023
right Shoulder BT	Math	-1.103	>0.999	-0.164
right Shoulder BT	Talk	-1.684	>0.999	-0.321
right Shoulder Vel. mean	Math	0.018	>0.999	0.004
right Shoulder Vel. mean	Talk	-0.316	>0.999	-0.059
right Shoulder SP	Math	-0.417	>0.999	-0.066
right Shoulder SP	Talk	-1.699	>0.999	-0.315
right Hand Visibility	Talk	-0.068	>0.999	-0.011
right Hand Visibility	Math	0.503	>0.999	0.094
head Vel. mean	Talk	1.650	>0.999	0.211
head BT	Talk	-2.016	>0.999	-0.318
left Shoulder Vel. std	Talk	0.298	>0.999	0.056
left Shoulder Vel. std	Math	0.030	>0.999	0.005
left Shoulder SP	Talk	1.486	>0.999	0.258
left Shoulder SP	Math	-0.978	>0.999	-0.188
left Shoulder Vel. mean	Talk	0.162	>0.999	0.028
left Shoulder Vel. mean	Math	0.246	>0.999	0.047
left Shoulder BT	Talk	1.618	>0.999	0.262
left Shoulder BT	Math	-1.058	>0.999	-0.189
left Hand Visibility	Talk	0.604	>0.999	0.104
left Hand Visibility	Math	1.274	>0.999	0.251
left Elbow Visibility	Talk	-0.408	>0.999	-0.060
right Shoulder Vel. std	Math	-0.082	>0.999	-0.016
right Shoulder Vel. std	Talk	-0.485	>0.999	-0.096

Table B.12: t-test results of gaze behavior features for the phases Math and Talk; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Feature	Phase	t(37)	p	Hedges' g
left Eye SP	Math	-3.527	0.222	-0.541
right Eye SP	Math	-2.974	0.992	-0.510
right Eye BT	Math	-2.848	>0.999	-0.471
left Eye BT	Math	-2.836	>0.999	-0.410
right Pupil Vel. mean	Math	0.011	>0.999	0.002
right Pupil std	Talk	-1.498	>0.999	-0.226
right Pupil std	Math	-1.537	>0.999	-0.269
right Pupil min	Talk	-2.201	>0.999	-0.303
right Pupil min	Math	-1.705	>0.999	-0.221
right BT	Talk	-1.257	>0.999	-0.188
right Pupil mean	Math	-1.843	>0.999	-0.261
right Pupil max	Talk	-2.332	>0.999	-0.316
right Pupil max	Math	-1.888	>0.999	-0.289
right Pupil diff	Talk	-1.632	>0.999	-0.319
right Eye Vel. mean	Talk	0.266	>0.999	0.034
right Pupil diff	Math	0.851	>0.999	0.174
right Pupil Blinks	Talk	1.408	>0.999	0.152
right Pupil Blinks	Math	-0.362	>0.999	-0.040
right Eye SP	Talk	-1.840	>0.999	-0.296
right Pupil mean	Talk	-2.182	>0.999	-0.310
left Eye Vel. std	Talk	-0.051	>0.999	-0.006
left Eye Vel. std	Math	-0.212	>0.999	-0.035
left BT	Talk	-1.025	>0.999	-0.141
left Pupil Blinks	Math	1.451	>0.999	0.198
left Pupil Blinks	Talk	1.102	>0.999	0.112
left Pupil diff	Math	1.005	>0.999	0.192
left Pupil diff	Talk	-1.273	>0.999	-0.276
left Pupil max	Math	-2.004	>0.999	-0.316
left Pupil max	Talk	-2.273	>0.999	-0.298
right Eye Vel. std	Math	-1.065	>0.999	-0.171

left Pupil mean	Math	-1.663	>0.999	-0.230
left Pupil min	Math	-1.147	>0.999	-0.159
left Pupil min	Talk	-2.070	>0.999	-0.280
left Pupil std	Math	-1.464	>0.999	-0.252
left Pupil std	Talk	-1.174	>0.999	-0.143
left SP	Talk	-1.919	>0.999	-0.306
left Pupil Vel. mean	Math	0.542	>0.999	0.080
left Pupil Vel. mean	Talk	0.102	>0.999	0.012
left Pupil mean	Talk	-2.146	>0.999	-0.283
right Eye Vel. std	Talk	0.049	>0.999	0.006

Appendix C

rPPG

```yaml
BASE: ['']
TOOLBOX_MODE: "only_test"
TEST:
  METRICS: ['MAE', 'RMSE', 'MAPE', 'Pearson']
  USE_LAST_EPOCH: True
  DATA:
    FS: 30
    DATASET: PURE
    DO_PREPROCESS: False
    DATA_FORMAT: NDCHW
    DATA_PATH: "../PURE/RawData"
    CACHED_PATH: "../PreprocessedData"
    PREPROCESS:
      DATA_TYPE: [ 'DiffNormalized','Standardized' ]
      LABEL_TYPE: DiffNormalized
      DO_CHUNK: True
      CHUNK_LENGTH: 180
      CROP_FACE:
        DO_CROP_FACE: True
        USE_LARGE_FACE_BOX: True
        LARGE_BOX_COEF: 1.5
        DETECTION:
          DO_DYNAMIC_DETECTION: False
          DYNAMIC_DETECTION_FREQUENCY : 30
          USE_MEDIAN_FACE_BOX: False
      RESIZE:
        H: 72
        W: 72
MODEL:
  DROP_RATE: 0.2
  NAME: Tscan
  TSCAN:
    FRAME_DEPTH: 10
INFERENCE:
  BATCH_SIZE: 4
  EVALUATION_METHOD: FFT
  EVALUATION_WINDOW:
    USE_SMALLER_WINDOW: False
    WINDOW_SIZE: 10       # In seconds
  MODEL_PATH:   "./final_model_release/UBFC-rPPG_TSCAN.pth"
```

Listing 1: YAML example file for the model TSCAN which was trained on UBFC-rPPG. It shows the configuration for an inference task on the PURE dataset.

List of Figures

List of Tables

Bibliography

[Abb21] Anzar Abbas, Colin Sauder, Vijay Yadav, Vidya Koesmahargyo, Allison Aghjayan, Serena Marecki, Miriam Evans, and Isaac R. Galatzer-Levy. "Remote Digital Measurement of Facial and Vocal Markers of Major Depressive Disorder Severity and Treatment Response: A Pilot Study". In: *Frontiers in Digital Health* 3 (Mar. 31, 2021), p. 610006. ISSN: 2673-253X. DOI: 10.3389/fdgth.2021.610006. URL: https://www.frontiersin.org/articles/10.3389/fdgth.2021.610006/full (visited on 04/21/2024).

[Aig18] Jonathan Aigrain, Michel Spodenkiewicz, Severine Dubuisson, Marcin Detyniecki, David Cohen, and Mohamed Chetouani. "Multimodal Stress Detection from Multiple Assessments". In: *IEEE Transactions on Affective Computing* 9.4 (Oct. 1, 2018), pp. 491–506. ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2016.2631594. URL: https://ieeexplore.ieee.org/document/7752842/ (visited on 01/15/2024).

[Al-17] Ali Al-Naji, Kim Gibson, Sang-Heon Lee, and Javaan Chahl. "Monitoring of Cardiorespiratory Signal: Principles of Remote Measurements and Review of Methods". In: *IEEE Access* 5 (2017), pp. 15776–15790. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2017.2735419. URL: http://ieeexplore.ieee.org/document/8002579/ (visited on 03/09/2024).

[All07] John Allen. "Photoplethysmography and its application in clinical physiological measurement". In: *Physiological Measurement* 28.3 (Mar. 1, 2007), R1–R39. ISSN: 0967-3334, 1361-6579. DOI: 10.1088/0967-3334/28/3/R01. URL: https://iopscience.iop.org/article/10.1088/0967-3334/28/3/R01 (visited on 02/06/2024).

[All17] Andrew P. Allen, Paul J. Kennedy, Samantha Dockray, John F. Cryan, Timothy G. Dinan, and Gerard Clarke. "The Trier Social Stress Test: Principles and practice". In: *Neurobiology of Stress* 6 (Feb. 2017), pp. 113–126. ISSN: 23522895. DOI: 10.1016/j.ynstr.2016.11.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352289516300224 (visited on 04/23/2024).

[And17] Gustavo Xavier Andrade Miranda. "Analyzing of the vocal fold dynamics using laryngeal videos". PhD thesis. Universidad Politécnica de Madrid, 2017. DOI: 10.20868/UPM.thesis.47122. URL: http://oa.upm.es/47122/ (visited on 03/03/2024).

[APA19] APA. "Stress in America 2019"". In: *Am. Psychol. Assoc.* (2019).

[Arn10] B. Arnrich, C. Setz, R. La Marca, G. Troster, and U. Ehlert. "What Does Your Chair Know About Your Stress Level?" In: *IEEE Transactions on Information Technology in Biomedicine* 14.2 (Mar. 2010), pp. 207–214. ISSN: 1089-7771. DOI: 10.1109/TITB.2009.2035498. URL: http://ieeexplore.ieee.org/document/5308331/ (visited on 03/05/2024).

[Bai19] Alice Baird, Shahin Amiriparian, Nicholas Cummins, Sarah Sturmbauer, Johanna Janson, Eva-Maria Messner, Harald Baumeister, Nicolas Rohleder, and Björn W. Schuller. "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test". In: *Interspeech 2019*. Interspeech 2019. ISCA, Sept. 15, 2019, pp. 534–538. DOI: 10.21437/Interspeech.2019-1352. URL: https://www.isca-archive.org/interspeech_2019/baird19_interspeech.html (visited on 01/20/2024).

[Bai21] Alice Baird, Andreas Triantafyllopoulos, Sandra Zänkert, Sandra Ottl, Lukas Christ, Lukas Stappen, Julian Konzok, Sarah Sturmbauer, Eva-Maria Meßner, Brigitte M. Kudielka, Nicolas Rohleder, Harald Baumeister, and Björn W. Schuller. "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress". In: *Frontiers in Computer Science* 3 (Dec. 6, 2021), p. 750284. ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.750284. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2021.750284/full (visited on 01/20/2024).

[Bal18] Tadas Baltru{\textbackslash}vsaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. "OpenFace 2.0: Facial Behavior Analysis Toolkit". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (2018), pp. 59–66. URL: https://api.semanticscholar.org/CorpusID:206652126.

[Bar07] Yair Bar-Haim, Dominique Lamy, Lee Pergamin, Marian J. Bakermans-Kranenburg, and Marinus H. Van IJzendoorn. "Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study." In: *Psychological Bulletin* 133.1 (2007), pp. 1–24. ISSN: 1939-1455, 0033-2909. DOI: 10.1037/0033-2909.133.1.1. URL: https://doi.apa.org/doi/10.1037/0033-2909.133.1.1 (visited on 03/05/2024).

[Bel04] Pascal Belin, Shirley Fecteau, and Catherine Bédard. "Thinking the voice: neural correlates of voice perception". In: *Trends in Cognitive Sciences* 8.3 (Mar. 2004), pp. 129–135. ISSN: 13646613. DOI: 10.1016/j.tics.2004.01.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661304000257 (visited on 03/05/2024).

[Ben19] Y. Benezeth, S. Bobbia, K. Nakamura, R. Gomez, and J. Dubois. "Probabilistic Signal Quality Metric for Reduced Complexity Unsupervised Remote Photoplethysmography". In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT). Oslo, Norway: IEEE, May 2019, pp. 1–5. ISBN: 978-1-72812-342-4. DOI: 10.1109/ISMICT.2019.8744004. URL: https://ieeexplore.ieee.org/document/8744004/ (visited on 02/12/2024).

[Bla23] Jost U. Blasberg, Mathilde Gallistl, Magdalena Degering, Felicitas Baierlein, and Veronika Engert. "You look stressed: A pilot study on facial action unit activity in the context of psychosocial stress". In: *Comprehensive Psychoneuroendocrinology* 15 (Aug. 2023), p. 100187. ISSN: 26664976. DOI: 10.1016/j.cpnec.2023.100187. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666497623000218 (visited on 01/14/2024).

[Bla86] J. M. Bland and D. G. Altman. "Statistical methods for assessing agreement between two methods of clinical measurement". In: *Lancet* 1.8476 (1986), pp. 307–10. ISSN: 0140-6736 (Print) 0140-6736.

[Bob01] A.F. Bobick and J.W. Davis. "The recognition of human movement using temporal templates". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.3 (Mar. 2001), pp. 257–267. ISSN: 01628828. DOI: 10.1109/34.910878. URL: http://ieeexplore.ieee.org/document/910878/ (visited on 03/09/2024).

[Bob19] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. "Unsupervised skin tissue segmentation for remote photoplethysmography". In: *Pattern Recognition Letters* 124 (June 2019), pp. 82–90. ISSN: 01678655. DOI: 10.1016/j.patrec.2017.10.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167865517303860 (visited on 02/13/2024).

[Bra08] Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. "The pupil as a measure of emotional arousal and autonomic activation". In: *Psychophysiology* 45.4 (July 2008), pp. 602–607. ISSN: 0048-5772, 1469-8986. DOI: 10.1111/j.1469-

8986.2008.00654.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8986.2008.00654.x (visited on 01/20/2024).

[Buc14] Tony W. Buchanan, Jacqueline S. Laures-Gore, and Melissa C. Duff. "Acute stress reduces speech fluency". In: *Biological Psychology* 97 (Mar. 2014), pp. 60–66. ISSN: 03010511. DOI: 10.1016/j.biopsycho.2014.02.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0301051114000441 (visited on 01/20/2024).

[Cao21] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (Jan. 1, 2021), pp. 172–186. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2019.2929257. URL: https://ieeexplore.ieee.org/document/8765346/ (visited on 04/23/2024).

[Car16] Roger Carpenter. "A Review of Instruments on Cognitive Appraisal of Stress". In: *Archives of Psychiatric Nursing* 30.2 (Apr. 2016), pp. 271–279. ISSN: 08839417. DOI: 10.1016/j.apnu.2015.07.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0883941715001429 (visited on 02/15/2024).

[Che18a] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. "MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices". In: (2018). Publisher: arXiv Version Number: 4. DOI: 10.48550/ARXIV.1804.07573. URL: https://arxiv.org/abs/1804.07573 (visited on 02/23/2024).

[Che18b] Weixuan Chen and Daniel McDuff. "DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks". In: (2018). Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1805.07888. URL: https://arxiv.org/abs/1805.07888 (visited on 02/13/2024).

[Che20] Zhaokang Chen and Bertram E. Shi. "Offset Calibration for Appearance-Based Gaze Estimation via Gaze Decomposition". In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass Village, CO, USA: IEEE, Mar. 2020, pp. 259–268. ISBN: 978-1-72816-553-0. DOI: 10.1109/WACV45572.2020.9093419. URL: https://ieeexplore.ieee.org/document/9093419/ (visited on 04/23/2024).

[Che21] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J. Chang. "Py-Feat: Python Facial Expression Analysis Toolbox". In: (2021). Publisher: arXiv Version Number: 4. DOI: 10.48550/ARXIV.2104.03509. URL: https://arxiv.org/abs/2104.03509 (visited on 02/23/2024).

[Chi21] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation". In: *PeerJ Computer Science* 7 (July 5, 2021), e623. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.623. URL: https://peerj.com/articles/cs-623 (visited on 03/26/2024).

[Cho24] Myounglee Choo, Doeun Park, Minseo Cho, Sujin Bae, Jinwoo Kim, and Doug Hyun Han. "Exploring a multimodal approach for utilizing digital biomarkers for childhood mental health screening". In: *Frontiers in Psychiatry* 15 (Apr. 11, 2024), p. 1348319. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2024.1348319. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2024.1348319/full (visited on 04/21/2024).

[Dar72] Charles Darwin. ""The expression of the emotions in man and animals."" In: *Chicago: University of Chicago Press.* (1872).

[Das21] Ananyananda Dasari, Sakthi Kumar Arul Prakash, László A. Jeni, and Conrad S. Tucker. "Evaluation of biases in remote photoplethysmography methods". In: *npj Digital Medicine* 4.1 (June 3, 2021), p. 91. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00462-z. URL: https://www.nature.com/articles/s41746-021-00462-z (visited on 02/04/2024).

[De 06] Ana Clara Naufel De Felippe, Maria Helena Marotti Martelletti Grillo, and Thaís Helena Grechi. "Standardization of acoustic measures for normal voice patterns". In: *Brazilian Journal of Otorhinolaryngology* 72.5 (Sept. 2006), pp. 659–664. ISSN: 18088694. DOI: 10.1016/S1808-8694(15)31023-5. URL: https://linkinghub.elsevier.com/retrieve/pii/S1808869415310235 (visited on 03/05/2024).

[De 13] Gerard De Haan and Vincent Jeanne. "Robust Pulse Rate From Chrominance-Based rPPG". In: *IEEE Transactions on Biomedical Engineering* 60.10 (Oct. 2013), pp. 2878–2886. ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2013.2266196. URL: https://ieeexplore.ieee.org/document/6523142/ (visited on 02/13/2024).

[De 14] G De Haan and A Van Leest. "Improved motion robustness of remote-PPG by using the blood volume pulse signature". In: *Physiological Measurement* 35.9 (Sept. 1, 2014), pp. 1913–1926. ISSN: 0967-3334, 1361-6579. DOI: 10.1088/0967-3334/35/9/1913. URL: https://iopscience.iop.org/article/10.1088/0967-3334/35/9/1913 (visited on 02/24/2024).

[Den19] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. *RetinaFace: Single-stage Dense Face Localisation in the Wild*. May 4, 2019. arXiv: 1905.00641[cs]. URL: http://arxiv.org/abs/1905.00641 (visited on 02/18/2024).

[Di 24] Daniele Di Lernia, Gianluca Finotti, Manos Tsakiris, Giuseppe Riva, and Marnix Naber. "Remote photoplethysmography (rPPG) in the wild: Remote heart rate imaging via online webcams". In: *Behavior Research Methods* (Apr. 17, 2024). ISSN: 1554-3528. DOI: 10.3758/s13428-024-02398-0. URL: https://link.springer.com/10.3758/s13428-024-02398-0 (visited on 04/24/2024).

[Dic04] Sally S. Dickerson and Margaret E. Kemeny. "Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research." In: *Psychological Bulletin* 130.3 (2004), pp. 355–391. ISSN: 1939-1455, 0033-2909. DOI: 10.1037/0033-2909.130.3.355. URL: https://doi.apa.org/doi/10.1037/0033-2909.130.3.355 (visited on 02/15/2024).

[Ekm03] P. Ekman and W.V. Friesen. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Spectrum book Bd. 10. Malor Books, 2003. ISBN: 978-1-883536-36-7. URL: https://books.google.com/books?id=TukNoJDgMTUC.

[Ekm78] Paul Ekman and Wallace V. Friesen. "Facial Action Coding System". In: *Environmental Psychology & Nonverbal Behavior* (1978).

[Emp24] EmpkinS. *EmpkinS – Website für den SFB-Antrag Empathokinästhetische Sensorik*. 2024. URL: https://empkins.de/ (visited on 02/05/2024).

[Epe18] Elissa S. Epel, Alexandra D. Crosswell, Stefanie E. Mayer, Aric A. Prather, George M. Slavich, Eli Puterman, and Wendy Berry Mendes. "More than a feeling: A unified view of stress measurement for population science". In: *Frontiers in Neuroendocrinology* 49 (Apr. 2018), pp. 146–169. ISSN: 00913022. DOI: 10.1016/j.yfrne.2018.03.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0091302218300219 (visited on 04/24/2024).

[Etc92] Nancy L. Etcoff and John J. Magee. "Categorical perception of facial expressions". In: *Cognition* 44.3 (Jan. 1992), pp. 227–240. ISSN: 00100277. DOI: 10.1016/0010-0277(92)90002-Y. URL: https://linkinghub.elsevier.com/retrieve/pii/001002779290002Y (visited on 03/05/2024).

[Fan23] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (June 1, 2023), pp. 7157–7173. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2022.3222784. URL: https://ieeexplore.ieee.org/document/9954214/ (visited on 04/23/2024).

[Foa86] Edna B. Foa and Michael J. Kozak. "Emotional processing of fear: Exposure to corrective information." In: *Psychological Bulletin* 99.1 (Jan. 1986), pp. 20–35. ISSN: 1939-1455, 0033-2909. DOI: 10.1037/0033-2909.99.1.20. URL: https://doi.apa.org/doi/10.1037/0033-2909.99.1.20 (visited on 03/05/2024).

[Fri07] Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. "Gaze cueing of attention: Visual attention, social cognition, and individual differences." In: *Psychological Bulletin* 133.4 (July 2007), pp. 694–724. ISSN: 1939-1455, 0033-2909. DOI: 10.1037/0033-2909.133.4.694. URL: https://doi.apa.org/doi/10.1037/0033-2909.133.4.694 (visited on 03/05/2024).

[Fuk17] Munenori Fukunishi, Kouki Kurita, Shoji Yamamoto, and Norimichi Tsumura. "Non-contact video-based estimation of heart rate variability spectrogram from hemoglobin composition". In: *Artificial Life and Robotics* 22.4 (Dec. 2017), pp. 457–463. ISSN: 1433-5298, 1614-7456. DOI: 10.1007/s10015-017-0382-1. URL: http://link.springer.com/10.1007/s10015-017-0382-1 (visited on 03/09/2024).

[Gaa09] Jens Gaab. "PASA – Primary Appraisal Secondary Appraisal". In: *Verhaltenstherapie* 19.2 (2009), pp. 114–115. ISSN: 1423-0402, 1016-6262. DOI: 10.1159/000223610. URL: https://www.karger.com/Article/FullText/223610 (visited on 02/15/2024).

[Gia12] Dimitris Giakoumis, Anastasios Drosou, Pietro Cipresso, Dimitrios Tzovaras, George Hassapis, Andrea Gaggioli, and Giuseppe Riva. "Using Activity-Related Behavioural Features towards More Effective Automatic Stress Detection". In: *PLoS ONE* 7.9 (Sept. 19, 2012). Ed. by Tiziana Zalla, e43571. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0043571. URL: https://dx.plos.org/10.1371/journal.pone.0043571 (visited on 01/21/2024).

[Gia17] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P.G. Simos, K. Marias, and M. Tsiknakis. "Stress and anxiety detection using facial cues from videos". In: *Biomedical Signal Processing and Control* 31 (Jan. 2017), pp. 89–101.

ISSN: 17468094. DOI: 10.1016/j.bspc.2016.06.020. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809416300805 (visited on 01/21/2024).

[Goo17] William K. Goodman, Johanna Janson, and Jutta M. Wolf. "Meta-analytical assessment of the effects of protocol variations on cortisol responses to the Trier Social Stress Test". In: *Psychoneuroendocrinology* 80 (June 2017), pp. 26–35. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2017.02.030. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453016309702 (visited on 04/23/2024).

[Guy23] Nitzan Guy, Hagar Azulay, Yoni Pertzov, and Salomon Israel. "Attenuation of visual exploration following stress". In: *Psychophysiology* 60.10 (Oct. 2023), e14330. ISSN: 0048-5772, 1469-8986. DOI: 10.1111/psyp.14330. URL: https://onlinelibrary.wiley.com/doi/10.1111/psyp.14330 (visited on 01/15/2024).

[Ham20] Ajna Hamidovic, Kristina Karapetyan, Fadila Serdarevic, So Hee Choi, Tory Eisenlohr-Moul, and Graziano Pinna. "Higher Circulating Cortisol in the Follicular vs. Luteal Phase of the Menstrual Cycle: A Meta-Analysis". In: *Frontiers in Endocrinology* 11 (June 2, 2020), p. 311. ISSN: 1664-2392. DOI: 10.3389/fendo.2020.00311. URL: https://www.frontiersin.org/article/10.3389/fendo.2020.00311/full (visited on 02/17/2024).

[Hap21] Johanna Happold, Robert Richer, Arne Kuderle, Heiko Gabner, Jochen Klucken, Bjoern M. Eskofier, and Felix Kluge. "Evaluation of Orthostatic Reactions in Real-World Environments Using Wearable Sensors". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Mexico: IEEE, Nov. 1, 2021, pp. 6987–6990. ISBN: 978-1-72811-179-7. DOI: 10.1109/EMBC46164.2021.9630842. URL: https://ieeexplore.ieee.org/document/9630842/ (visited on 02/23/2024).

[Her17] Nadja Herten, Tobias Otto, and Oliver T. Wolf. "The role of eye fixation in memory enhancement under stress – An eye tracking study". In: *Neurobiology of Learning and Memory* 140 (Apr. 2017), pp. 134–144. ISSN: 10747427. DOI: 10.1016/j.nlm.2017.02.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S1074742716302325 (visited on 01/15/2024).

[Het09] S. Het, N. Rohleder, D. Schoofs, C. Kirschbaum, and O.T. Wolf. "Neuroendocrine and psychometric evaluation of a placebo version of the 'Trier Social Stress Test'". In: *Psychoneuroendocrinology* 34.7 (Aug. 2009), pp. 1075–1086. ISSN: 03064530. DOI:

10.1016/j.psyneuen.2009.02.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453009000614 (visited on 02/15/2024).

[Heu17] Guillaume Heusch, André Anjos, and Sébastien Marcel. "A Reproducible Study on Remote Heart Rate Measurement". In: (2017). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1709.00962. URL: https://arxiv.org/abs/1709.00962 (visited on 02/13/2024).

[Hon08] Kiyoshi Honda. "Physiological Processes of Speech Production". In: *Springer Handbook of Speech Processing*. Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Arden Huang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 7–26. ISBN: 978-3-540-49125-5 978-3-540-49127-9. DOI: 10.1007/978-3-540-49127-9_2. URL: http://link.springer.com/10.1007/978-3-540-49127-9_2 (visited on 03/05/2024).

[Ins17] Thomas R. Insel. "Digital Phenotyping: Technology for a New Science of Behavior". In: *JAMA* 318.13 (Oct. 3, 2017), p. 1215. ISSN: 0098-7484. DOI: 10.1001/jama.2017.11295. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.11295 (visited on 02/06/2024).

[Iuc20] Kaito Iuchi, Ryota Mitsuhashi, Takashi Goto, Akira Matsubara, Takahiro Hirayama, Hideki Hashizume, and Norimichi Tsumura. "Stress levels estimation from facial video based on non-contact measurement of pulse wave". In: *Artificial Life and Robotics* 25.3 (Aug. 2020), pp. 335–342. ISSN: 1433-5298, 1614-7456. DOI: 10.1007/s10015-020-00624-4. URL: https://link.springer.com/10.1007/s10015-020-00624-4 (visited on 02/11/2024).

[Kah66] Daniel Kahneman and Jackson Beatty. "Pupil Diameter and Load on Memory". In: *Science* 154.3756 (Dec. 23, 1966), pp. 1583–1585. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.154.3756.1583. URL: https://www.science.org/doi/10.1126/science.154.3756.1583 (visited on 03/05/2024).

[Kaz79] Alan E. Kazdin. "UNOBTRUSIVE MEASURES IN BEHAVIORAL ASSESSMENT". In: *Journal of Applied Behavior Analysis* 12.4 (Dec. 1979), pp. 713–724. ISSN: 0021-8855, 1938-3703. DOI: 10.1901/jaba.1979.12-713. URL: https://onlinelibrary.wiley.com/doi/10.1901/jaba.1979.12-713 (visited on 02/06/2024).

[Kir11] Christin Kirchhübel, David M. Howard, and Alex W. Stedmon. "Acoustic Correlates of Speech when Under Stress: Research, Methods and Future Directions". In: *International Journal of Speech, Language and the Law* 18.1 (Sept. 13, 2011), pp. 75–98.

ISSN: 1748-8893, 1748-8885. DOI: 10.1558/ijsll.v18i1.75. URL: https://journal.equinoxpub.com/IJSLL/article/view/5979 (visited on 01/20/2024).

[Kir93] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. "The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting". In: *Neuropsychobiology* 28.1 (1993), pp. 76–81. ISSN: 0302-282X, 1423-0224. DOI: 10.1159/000119004. URL: https://www.karger.com/Article/FullText/119004 (visited on 02/06/2024).

[Kun19] Kunaratnam Kunarakulan and Ahamed Rameez Mohamed Nizzad. "Real-Time Feeling Detection through Facial Expression Recognition: A Machine Learning Approach". In: Dec. 2019.

[Kwa17] Sang Gyu Kwak and Jong Hae Kim. "Central limit theorem: the cornerstone of modern statistics". In: *Korean Journal of Anesthesiology* 70.2 (2017), p. 144. ISSN: 2005-6419, 2005-7563. DOI: 10.4097/kjae.2017.70.2.144. URL: http://ekja.org/journal/view.php?doi=10.4097/kjae.2017.70.2.144 (visited on 02/23/2024).

[Lab19] Izelle Labuschagne, Caitlin Grace, Peter Rendell, Gill Terrett, and Markus Heinrichs. "An introductory guide to conducting the Trier Social Stress Test". In: *Neuroscience & Biobehavioral Reviews* 107 (Dec. 2019), pp. 686–695. ISSN: 01497634. DOI: 10.1016/j.neubiorev.2019.09.032. URL: https://linkinghub.elsevier.com/retrieve/pii/S0149763419308255 (visited on 04/23/2024).

[Las20] J. Lasselin, T. Sundelin, P.M. Wayne, M.J. Olsson, S. Paues Göranson, J. Axelsson, and M. Lekander. "Biological motion during inflammation in humans". In: *Brain, Behavior, and Immunity* 84 (Feb. 2020), pp. 147–153. ISSN: 08891591. DOI: 10.1016/j.bbi.2019.11.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0889159119306427 (visited on 02/23/2024).

[Laz84] Richard S. Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[Lew11] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak. "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity". In: *2011 federated conference on computer science and information systems (FedCSIS)* (2011), pp. 405–410.

[Lim21] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting". In: *International Journal of Forecasting* 37.4 (Oct. 2021), pp. 1748–1764. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2021.03.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207021000637 (visited on 05/06/2024).

[Liu21a] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. *Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement*. Feb. 28, 2021. arXiv: 2006.03790[cs,eess]. URL: http://arxiv.org/abs/2006.03790 (visited on 02/13/2024).

[Liu21b] Xin Liu, Brian L. Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. "EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Vitals Measurement". In: (2021). Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.2110.04447. URL: https://arxiv.org/abs/2110.04447 (visited on 02/13/2024).

[Liu23] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Soumyadip Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. *rPPG-Toolbox: Deep Remote PPG Toolbox*. Nov. 24, 2023. arXiv: 2210.00716[cs]. URL: http://arxiv.org/abs/2210.00716 (visited on 02/10/2024).

[Lug19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. "MediaPipe: A Framework for Building Perception Pipelines". In: (2019). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1906.08172. URL: https://arxiv.org/abs/1906.08172 (visited on 02/23/2024).

[Lun17] Scott Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: (2017). Publisher: [object Object] Version Number: 2. DOI: 10.48550/ARXIV.1705.07874. URL: https://arxiv.org/abs/1705.07874 (visited on 03/25/2024).

[Lun20] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1 (Jan. 17, 2020), pp. 56–67. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0138-9. URL: https://www.nature.com/articles/s42256-019-0138-9 (visited on 03/25/2024).

[Lup14] Sarah B. Lupis, Michelle Lerman, and Jutta M. Wolf. "Anger responses to psychosocial stress predict heart rate and cortisol stress responses in men but not women". In: *Psychoneuroendocrinology* 49 (Nov. 2014), pp. 84–95. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2014.07.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453014002558 (visited on 01/14/2024).

[Mak21] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. "NeuroKit2: A Python toolbox for neurophysiological signal processing". In: *Behavior Research Methods* 53.4 (Aug. 2021), pp. 1689–1696. ISSN: 1554-3528. DOI: 10.3758/s13428-020-01516-y. URL: https://link.springer.com/10.3758/s13428-020-01516-y (visited on 02/23/2024).

[Mal96] Marek Malik. "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use". In: *Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology* 93 (1996), pp. 1043–1065.

[McD14] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. "Remote measurement of cognitive stress via heart rate variability". In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Chicago, IL: IEEE, Aug. 2014, pp. 2957–2960. ISBN: 978-1-4244-7929-0. DOI: 10.1109/EMBC.2014.6944243. URL: http://ieeexplore.ieee.org/document/6944243/ (visited on 02/09/2024).

[McD15] Daniel J. McDuff, Justin R. Estepp, Alyssa M. Piasecki, and Ethan B. Blackford. "A survey of remote optical photoplethysmographic imaging methods". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Milan: IEEE, Aug. 2015, pp. 6398–6404. ISBN: 978-1-4244-9271-8. DOI: 10.1109/EMBC.2015.7319857. URL: http://ieeexplore.ieee.org/document/7319857/ (visited on 02/04/2024).

[McD17] Daniel J. McDuff, Ethan B. Blackford, and Justin R. Estepp. "The Impact of Video Compression on Remote Cardiac Pulse Measurement Using Imaging Photoplethysmography". In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. 2017 12th IEEE International Conference on Au-

tomatic Face & Gesture Recognition (FG 2017). Washington, DC, DC, USA: IEEE, May 2017, pp. 63–70. ISBN: 978-1-5090-4023-0. DOI: 10.1109/FG.2017.17. URL: http://ieeexplore.ieee.org/document/7961724/ (visited on 04/24/2024).

[Mor22] Hector Manuel Morales-Fajardo, Jorge Rodríguez-Arce, Alejandro Gutiérrez-Cedeño, José Caballero Viñas, José Javier Reyes-Lagos, Eric Alonso Abarca-Castro, Claudia Ivette Ledesma-Ramírez, and Adriana H. Vilchis-González. "Towards a Non-Contact Method for Identifying Stress Using Remote Photoplethysmography in Academic Environments". In: *Sensors* 22.10 (May 16, 2022), p. 3780. ISSN: 1424-8220. DOI: 10.3390/s22103780. URL: https://www.mdpi.com/1424-8220/22/10/3780 (visited on 02/09/2024).

[Nat09] U.M. Nater and N. Rohleder. "Salivary alpha-amylase as a non-invasive biomarker for the sympathetic nervous system: Current state of research". In: *Psychoneuroendocrinology* 34.4 (May 2009), pp. 486–496. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2009.01.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453009000328 (visited on 02/29/2024).

[Ni21] Aoxin Ni, Arian Azarang, and Nasser Kehtarnavaz. "A Review of Deep Learning-Based Contactless Heart Rate Measurement Methods". In: *Sensors* 21.11 (May 27, 2021), p. 3719. ISSN: 1424-8220. DOI: 10.3390/s21113719. URL: https://www.mdpi.com/1424-8220/21/11/3719 (visited on 02/10/2024).

[Nor22a] Matthias Norden, Amin Gerard Hofmann, Martin Meier, Felix Balzer, Oliver T Wolf, Erwin Böttinger, and Hanna Drimalla. "Inducing and Recording Acute Stress Responses on a Large Scale With the Digital Stress Test (DST): Development and Evaluation Study". In: *Journal of Medical Internet Research* 24.7 (July 15, 2022), e32280. ISSN: 1438-8871. DOI: 10.2196/32280. URL: https://www.jmir.org/2022/7/e32280 (visited on 05/06/2024).

[Nor22b] Matthias Norden, Oliver T. Wolf, Lennart Lehmann, Katja Langer, Christoph Lippert, and Hanna Drimalla. "Automatic Detection of Subjective, Annotated and Physiological Stress Responses from Video Data". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII). Nara, Japan: IEEE, Oct. 18, 2022, pp. 1–8. ISBN: 978-1-66545-908-2. DOI: 10.1109/ACII55700.2022.9953894. URL: https://ieeexplore.ieee.org/document/9953894/ (visited on 01/16/2024).

[OCo21] Daryl B. O'Connor, Julian F. Thayer, and Kavita Vedhara. "Stress and Health: A Review of Psychobiological Processes". In: *Annual Review of Psychology* 72.1 (Jan. 4, 2021), pp. 663–688. ISSN: 0066-4308, 1545-2085. DOI: 10.1146/annurev-psych-062520-122331. URL: https://www.annualreviews.org/doi/10.1146/annurev-psych-062520-122331 (visited on 02/06/2024).

[Oes23] Marie Oesten, Robert Richer, Luca Abel, Nicolas Rohleder, and Bjoern M. Eskofier. "VoStress – Voice-based Detection of Acute Psychosocial Stress". In: *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). Pittsburgh, PA, USA: IEEE, Oct. 15, 2023, pp. 1–4. DOI: 10.1109/BHI58575.2023.10313458. URL: https://ieeexplore.ieee.org/document/10313458/ (visited on 01/20/2024).

[Ogr19] Marissa Ogren, Brianna Kaplan, Yujia Peng, Kerri L. Johnson, and Scott P. Johnson. "Motion or emotion: Infants discriminate emotional biological motion based on low-level visual information". In: *Infant Behavior and Development* 57 (Nov. 2019), p. 101324. ISSN: 01636383. DOI: 10.1016/j.infbeh.2019.04.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0163638319300116 (visited on 03/05/2024).

[Pha20] Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. "Problems and opportunities in training deep learning software systems: an analysis of variance". In: *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. ASE '20: 35th IEEE/ACM International Conference on Automated Software Engineering. Virtual Event Australia: ACM, Dec. 21, 2020, pp. 771–783. ISBN: 978-1-4503-6768-4. DOI: 10.1145/3324884.3416545. URL: https://dl.acm.org/doi/10.1145/3324884.3416545 (visited on 04/24/2024).

[Pha21] Luan Pham, The Huynh Vu, and Tuan Anh Tran. "Facial Expression Recognition Using Residual Masking Network". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, Jan. 10, 2021, pp. 4513–4519. ISBN: 978-1-72818-808-9. DOI: 10.1109/ICPR48806.2021.9411919. URL: https://ieeexplore.ieee.org/document/9411919/ (visited on 02/23/2024).

[Pil18] Christian S. Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. "Local Group Invariance for Heart Rate Estimation From Face Videos in the Wild". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2018.

[Pis18] Katarzyna Pisanski, Aleksander Kobylarek, Luba Jakubowska, Judyta Nowak, Amelia Walter, Kamil Błaszczyński, Magda Kasprzyk, Krystyna Łysenko, Irmina Sukiennik, Katarzyna Piątek, Tomasz Frackowiak, and Piotr Sorokowski. "Multimodal stress detection: Testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test". In: *Hormones and Behavior* 106 (Nov. 2018), pp. 52–61. ISSN: 0018506X. DOI: 10.1016/j.yhbeh.2018.08.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0018506X18301272 (visited on 01/20/2024).

[Poh10] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation". In: *Optics Express* 18.10 (May 10, 2010), p. 10762. DOI: 10.1364/OE.18.010762. URL: https://opg.optica.org/abstract.cfm?URI=oe-18-10-10762 (visited on 02/12/2024).

[Pol21] I. Polanowski. "Work-related stress, anxiety or depression statistics in Great Britain, 2021,"" in: *UK Health Saf.* (2021).

[Pru03] Jens C. Pruessner, Clemens Kirschbaum, Gunther Meinlschmid, and Dirk H Hellhammer. "Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change". In: *Psychoneuroendocrinology* 28.7 (Oct. 2003), pp. 916–931. ISSN: 03064530. DOI: 10.1016/S0306-4530(02)00108-7. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453002001087 (visited on 02/23/2024).

[Ric21a] Robert Richer, Arne Kuderle, Jana Dorr, Nicolas Rohleder, and Bjoern M. Eskofier. "Assessing the Influence of the Inner Clock on the Cortisol Awakening Response and Pre-Awakening Movement". In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). Athens, Greece: IEEE, July 27, 2021, pp. 1–4. ISBN: 978-1-66540-358-0. DOI: 10.1109/BHI50953.2021.9508529. URL: https://ieeexplore.ieee.org/document/9508529/ (visited on 02/17/2024).

[Ric21b] Robert Richer, Arne Küderle, Martin Ullrich, Nicolas Rohleder, and Bjoern Eskofier. "BioPsyKit: A Python package for the analysis of biopsychological data". In: *Journal*

of Open Source Software 6.66 (Oct. 12, 2021), p. 3702. ISSN: 2475-9066. DOI: 10.21105/joss.03702. URL: https://joss.theoj.org/papers/10.21105/joss.03702 (visited on 02/23/2024).

[Ric22] Robert Richer, Veronika Koch, Arne Kuederle, Victoria M"uller, Vanessa Wirth, Marc Stamminger, Nicolas Rohleder, and Bjoern Eskofier. "Detection of Acute Psychosocial Stress from Body Movements using Machine Learning". In: *Psychosomatic Medicine*. Ed. by Lippincott Williams Wilkins. Mar. 23, 2022, A55–A55.

[Ric24a] Robert Richer, Veronika Koch, Luca Abel, Felicitas Hauck, Miriam Kurz, Veronika Ringgold, Victoria Müller, Arne Küderle, Lena Schindler-Gmelch, Bjoern M. Eskofier, and Nicolas Rohleder. "Machine learning-based detection of acute psychosocial stress from body posture and movements". In: *Scientific Reports* 14.1 (Apr. 8, 2024), p. 8251. ISSN: 2045-2322. DOI: 10.1038/s41598-024-59043-1. URL: https://www.nature.com/articles/s41598-024-59043-1 (visited on 04/23/2024).

[Ric24b] Robert Richer, Ekaterina Varkentin, Kurmanzhan Kurmanbekova, Victoria Müller, Luca Abel, Veronika Ringgold, Arne Küderle, Irina Brich, Nicolas Rohleder, and Bjoern M. Eskofier. "- Stress+ – Towards an Open-Source Web Application for the Remote Induction of Acute Psychosocial Stress". In: *Psychoneuroendocrinology* 160 (Feb. 2024), p. 106870. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2023.106870. URL: https://linkinghub.elsevier.com/retrieve/pii/S030645302300848X (visited on 05/06/2024).

[Roe10] Karin Roelofs, Muriel A. Hagenaars, and John Stins. "Facing Freeze: Social Threat Induces Bodily Freeze in Humans". In: *Psychological Science* 21.11 (Nov. 2010), pp. 1575–1581. ISSN: 0956-7976, 1467-9280. DOI: 10.1177/0956797610384746. URL: http://journals.sagepub.com/doi/10.1177/0956797610384746 (visited on 02/04/2024).

[Roh19] Nicolas Rohleder. "Stress and inflammation – The need to address the gap in the transition between acute and chronic stress effects". In: *Psychoneuroendocrinology* 105 (July 2019), pp. 164–171. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2019.02.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453018306954 (visited on 05/07/2024).

[Sab23] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, and Fan Yang. "UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress". In: *IEEE Transactions on Affective Computing* 14.1 (Jan. 1, 2023),

pp. 622–636. ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2021.3056960. URL: https://ieeexplore.ieee.org/document/9346017/ (visited on 02/04/2024).

[Sch22a] K. Schultebraucks and B. Chang. "Digital biomarkers for predicting PTSD, depression, and burnout in emergency department clinicians". In: *Psychosomatic Medicine* (2022), A84–A84.

[Sch22b] Katharina Schultebraucks, Vijay Yadav, Arieh Y. Shalev, George A. Bonanno, and Isaac R. Galatzer-Levy. "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood". In: *Psychological Medicine* 52.5 (Apr. 2022), pp. 957–967. ISSN: 0033-2917, 1469-8978. DOI: 10.1017/S0033291720002718. URL: https://www.cambridge.org/core/product/identifier/S0033291720002718/type/journal_article (visited on 02/06/2024).

[Sch24] Katharina Schultebraucks, Zain Khan, Joseph Chang, and Bernard Chang. "Digital biomarkers for diagnostic assessment of stress pathologies and neurocognitive performance". In: *Psychoneuroendocrinology* 160 (Feb. 2024), p. 106754. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2023.106754. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453023007321 (visited on 04/21/2024).

[Sha13] Miraj Shah, David G. Cooper, Houwei Cao, Ruben C. Gur, Ani Nenkova, and Ragini Verma. "Action Unit Models of Facial Expression of Emotion in the Presence of Speech". In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII). Geneva, Switzerland: IEEE, Sept. 2013, pp. 49–54. ISBN: 978-0-7695-5048-0. DOI: 10.1109/ACII.2013.15. URL: http://ieeexplore.ieee.org/document/6681406/ (visited on 04/23/2024).

[Sla15] Danica C. Slavish, Jennifer E. Graham-Engeland, Joshua M. Smyth, and Christopher G. Engeland. "Salivary markers of inflammation in response to acute stress". In: *Brain, Behavior, and Immunity* 44 (Feb. 2015), pp. 253–269. ISSN: 08891591. DOI: 10.1016/j.bbi.2014.08.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0889159114004255 (visited on 02/06/2024).

[Spe18] R. Spetlik, V. Franc, and J. Matas. "Visual heart rate estimation with convolutional neural network". In: *Proceedings of the British Machine Vision Conference, Newcastle, UK* (2018), pp. 3–6.

[Str14] Ronny Stricker, Steffen Muller, and Horst-Michael Gross. "Non-contact video-based pulse rate measurement on a mobile service robot". In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 2014 RO-MAN: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. Edinburgh, UK: IEEE, Aug. 2014, pp. 1056–1062. DOI: 10.1109/ROMAN.2014.6926392. URL: http://ieeexplore.ieee.org/document/6926392/ (visited on 02/13/2024).

[Tay00] Shelley E. Taylor, Laura Cousino Klein, Brian P. Lewis, Tara L. Gruenewald, Regan A. R. Gurung, and John A. Updegraff. "Biobehavioral responses to stress in females: Tend-and-befriend, not fight-or-flight." In: *Psychological Review* 107.3 (2000), pp. 411–429. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.107.3.411. URL: https://doi.apa.org/doi/10.1037/0033-295X.107.3.411 (visited on 03/09/2024).

[Thi10] Tuan Hue Thi, Jian Zhang, Li Cheng, Li Wang, and Shinichi Satoh. "Human Action Recognition and Localization in Video Using Structured Learning of Local Space-Time Features". In: *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Boston, MA, USA: IEEE, Aug. 2010, pp. 204–211. ISBN: 978-1-4244-8310-5. DOI: 10.1109/AVSS.2010.76. URL: http://ieeexplore.ieee.org/document/5597147/ (visited on 04/23/2024).

[Ulr09] Yvonne M. Ulrich-Lai and James P. Herman. "Neural regulation of endocrine and autonomic stress responses". In: *Nature Reviews Neuroscience* 10.6 (June 2009), pp. 397–409. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn2647. URL: https://www.nature.com/articles/nrn2647 (visited on 02/06/2024).

[Vab19] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. "Machine learning algorithm validation with a limited sample size". In: *PLOS ONE* 14.11 (Nov. 7, 2019). Ed. by Enrique Hernandez-Lemus, e0224365. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0224365. URL: https://dx.plos.org/10.1371/journal.pone.0224365 (visited on 04/25/2024).

[Val18] Raphael Vallat. "Pingouin: statistics in Python". In: *Journal of Open Source Software* 3.31 (Nov. 19, 2018), p. 1026. ISSN: 2475-9066. DOI: 10.21105/joss.01026. URL: http://joss.theoj.org/papers/10.21105/joss.01026 (visited on 02/23/2024).

[Van07] Jan Van Den Stock, Ruthger Righart, and Beatrice De Gelder. "Body expressions influence recognition of emotions in the face and voice." In: *Emotion* 7.3 (Aug. 2007),

pp. 487–494. ISSN: 1931-1516, 1528-3542. DOI: 10.1037/1528-3542.7.3.487. URL: https://doi.apa.org/doi/10.1037/1528-3542.7.3.487 (visited on 03/05/2024).

[Van18] Martine Van Puyvelde, Xavier Neyt, Francis McGlone, and Nathalie Pattyn. "Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance". In: *Frontiers in Psychology* 9 (Nov. 20, 2018), p. 1994. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01994. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2018.01994/full (visited on 01/20/2024).

[Vat21] C. Carolyn Vatheuer, Antonia Vehlen, Bernadette Von Dawans, and Gregor Domes. "Gaze behavior is associated with the cortisol response to acute psychosocial stress in the virtual TSST". In: *Journal of Neural Transmission* 128.9 (Sept. 2021), pp. 1269–1278. ISSN: 0300-9564, 1435-1463. DOI: 10.1007/s00702-021-02344-w. URL: https://link.springer.com/10.1007/s00702-021-02344-w (visited on 01/14/2024).

[Ver08] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. "Remote plethysmographic imaging using ambient light". In: *Optics Express* 16.26 (Dec. 22, 2008), p. 21434. ISSN: 1094-4087. DOI: 10.1364/OE.16.021434. URL: https://opg.optica.org/abstract.cfm?URI=oe-16-26-21434 (visited on 02/09/2024).

[Vie18] Carla Viegas, Shing-Hon Lau, Roy Maxion, and Alexander Hauptmann. "Towards Independent Stress Detection: A Dependent Model Using Facial Action Units". In: *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. 2018 International Conference on Content-Based Multimedia Indexing (CBMI). La Rochelle: IEEE, Sept. 2018, pp. 1–6. ISBN: 978-1-5386-7021-7. DOI: 10.1109/CBMI.2018.8516497. URL: https://ieeexplore.ieee.org/document/8516497/ (visited on 01/15/2024).

[Vio01] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Vol. 1. Kauai, HI, USA: IEEE Comput. Soc, 2001, pp. I–511–I–518. ISBN: 978-0-7695-1272-3. DOI: 10.1109/CVPR.2001.990517. URL: http://ieeexplore.ieee.org/document/990517/ (visited on 03/26/2024).

[Wal86] Harald G. Wallbott and Klaus R. Scherer. "Cues and channels in emotion recognition." In: *Journal of Personality and Social Psychology* 51.4 (Oct. 1986), pp. 690–699. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.51.4.690. URL: https://doi.apa.org/doi/10.1037/0022-3514.51.4.690 (visited on 03/05/2024).

[Wan16] Wenjin Wang, Sander Stuijk, and Gerard De Haan. "A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation". In: *IEEE Transactions on Biomedical Engineering* 63.9 (Sept. 2016), pp. 1974–1984. ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2015.2508602. URL: https://ieeexplore.ieee.org/document/7355301/ (visited on 03/09/2024).

[Wan17] Wenjin Wang, Albertus C. Den Brinker, Sander Stuijk, and Gerard De Haan. "Algorithmic Principles of Remote PPG". In: *IEEE Transactions on Biomedical Engineering* 64.7 (July 2017), pp. 1479–1491. ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2016.2609282. URL: http://ieeexplore.ieee.org/document/7565547/ (visited on 02/10/2024).

[Wan18] Chin-An Wang, Talia Baird, Jeff Huang, Jonathan D. Coutinho, Donald C. Brien, and Douglas P. Munoz. "Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional Face Task". In: *Frontiers in Neurology* 9 (Dec. 3, 2018), p. 1029. ISSN: 1664-2295. DOI: 10.3389/fneur.2018.01029. URL: https://www.frontiersin.org/article/10.3389/fneur.2018.01029/full (visited on 01/20/2024).

[Wat88] David Watson, Lee Anna Clark, and Auke Tellegen. "Development and validation of brief measures of positive and negative affect: The PANAS scales." In: *Journal of Personality and Social Psychology* 54.6 (1988), pp. 1063–1070. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.54.6.1063. URL: https://doi.apa.org/doi/10.1037/0022-3514.54.6.1063 (visited on 02/17/2024).

[Wel91] Emo Welzl. "Smallest enclosing disks (balls and ellipsoids)". In: *New Results and New Trends in Computer Science*. Ed. by Hermann Maurer. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991, pp. 359–370. ISBN: 978-3-540-46457-0.

[Wie13] Uta S. Wiemers, Daniela Schoofs, and Oliver T. Wolf. "A friendly version of the Trier Social Stress Test does not activate the HPA axis in healthy men and women". In: *Stress* 16.2 (Mar. 2013), pp. 254–260. ISSN: 1025-3890, 1607-8888. DOI: 10.3109/10253890.2012.714427. URL: http://www.tandfonline.com/doi/full/10.3109/10253890.2012.714427 (visited on 02/06/2024).

[Xia24] Hanguang Xiao, Tianqi Liu, Yisha Sun, Yulin Li, Shiyi Zhao, and Alberto Avolio. "Remote photoplethysmography for heart rate measurement: A review". In: *Biomedical Signal Processing and Control* 88 (Feb. 2024), p. 105608. ISSN: 17468094. DOI: 10.1016/j.bspc.2023.105608. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809423010418 (visited on 02/09/2024).

[Yan22] Ze Yang, Haofei Wang, and Feng Lu. "Assessment of Deep Learning-Based Heart Rate Estimation Using Remote Photoplethysmography Under Different Illuminations". In: *IEEE Transactions on Human-Machine Systems* 52.6 (Dec. 2022), pp. 1236–1246. ISSN: 2168-2291, 2168-2305. DOI: 10.1109/THMS.2022.3207755. URL: https://ieeexplore.ieee.org/document/9913818/ (visited on 02/04/2024).

[Yu19] Zitong Yu, Xiaobai Li, and Guoying Zhao. "Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks". In: (2019). Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1905.02419. URL: https://arxiv.org/abs/1905.02419 (visited on 02/13/2024).

[Yu22] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. "PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, June 2022, pp. 4176–4186. ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.00415. URL: https://ieeexplore.ieee.org/document/9879451/ (visited on 02/13/2024).

[Zän20] Sandra Zänkert, Brigitte M. Kudielka, and Stefan Wüst. "Effect of sugar administration on cortisol responses to acute psychosocial stress". In: *Psychoneuroendocrinology* 115 (May 2020), p. 104607. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2020.104607. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453020300263 (visited on 02/17/2024).

[Zha14] Li Zhang, Alamgir Hossain, and Ming Jiang. "Intelligent Facial Action and emotion recognition for humanoid robots". In: *2014 International Joint Conference on Neural Networks (IJCNN)*. 2014 International Joint Conference on Neural Networks (IJCNN). Beijing, China: IEEE, July 2014, pp. 739–746. ISBN: 978-1-4799-1484-5 978-1-4799-6627-1. DOI: 10.1109/IJCNN.2014.6889647. URL: https://ieeexplore.ieee.org/document/6889647 (visited on 03/05/2024).

[Zha20] Qi Zhan, Wenjin Wang, and Gerard De Haan. "Analysis of CNN-based remote-PPG to understand limitations and sensitivities". In: *Biomedical Optics Express* 11.3 (Mar. 1, 2020), p. 1268. ISSN: 2156-7085, 2156-7085. DOI: 10.1364/BOE.382637. URL: https://opg.optica.org/abstract.cfm?URI=boe-11-3-1268 (visited on 02/13/2024).

[Zhe24] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. "Deep Learning-based Human Pose Estimation: A

Survey". In: *ACM Computing Surveys* 56.1 (Jan. 31, 2024), pp. 1–37. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3603618. URL: https://dl.acm.org/doi/10.1145/3603618 (visited on 04/23/2024).

[Zit19] Giuseppe Angelo Zito, Kallia Apazoglou, Anisoara Paraschiv-Ionescu, Kamiar Aminian, and Selma Aybek. "Abnormal postural behavior in patients with functional movement disorders during exposure to stress". In: *Psychoneuroendocrinology* 101 (Mar. 2019), pp. 232–239. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2018.11.020. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306453018311314 (visited on 03/05/2024).

Appendix D

Acronyms

TSST Trier Social Stress Test

f-TSST friendly TSST

ML Machine Learning

HR heart rate

HRV heart rate variability

rPPG remote photoplethysmography

PPG photoplethysmography

FACS Facial Action Coding System

AU Action Unit

IMUs inertial measurement units

PCA Principal Component Analysis

ICA Independent Component Analysis

CHROM chrominance model

POS plane-orthogonal-to-skin model

BVP blood volume pulse

TSM Temporal Shift Module

bpm beats per minute

SBMLR stepwise backward multiple regression

SNS sympathetic nervous system

HPA hypothalamic-pituitary-adrenocortical

fps frames per second

DRM dichromatic reflection model

RGB red-green-blue

DRM Dynamic Region Model

MTCCN Multi-task Cascaded Convolutional Networks

ROI Region of Interest

CNN Convolutional Neural Network

DL deep learning

EDA electrodermal activity

FFT fast Fourier transform

MAE Mean absolute error

MAPE Mean absolute percentage error

SNR signal-noise ratio

ρ Pearson Correlation

RMSE root mean squared error

sAA salivary alpha-amylase

rPPG-HR rPPG-derived HR

PANAS Positive and Negative Affect Schedule

MoCap Motion Capture Suits

STROOP Stroop Color and Word Test

STAI State-Trait Anxiety Inventory

PASA Primary Appraisal Secondary Appraisal

ECG Electrocardiogram

IMU Inertial Measurement Unit

RR R-peak-to-R-peak

STADI State-Trait Anxiety-Depression Inventory

ANOVA Analysis of Variance

SHAP SHapley Additive exPlanation

RFE Recursive Feature Elimination

SkB Select-k-Best

NB Naïve Bayes

kNN k-Nearest-Neighbors

DT Decision Tree

SVM Support Vector Machine

Ada AdaBoost

RF Random Forest

CV cross-validation

HOG Histogram of Oriented Gradients

BMI Body Mass Index

SVR Support Vector Regressor

LGI Local Group Invariance