technische universität
dortmund

Impact of Varying Collision Avoidance Strategies on Human Stress Level in Human-Robot Interaction

Master Thesis

submitted to

Institute of Control Theory and Systems Engineering

Faculty of Electrical Engineering and Information Technology

Technische Universität Dortmund

by

Mohammed Faizan

Chennai, India

Date of Submission: February 19, 2024

Responsible Professor:
Univ.-Prof. Dr.-Ing. Prof. h.c. Dr. h.c. Torsten Bertram

Academic Supervisors:
M.Sc. Heiko Renz
M.Sc. Khazar Dargahi Nobari

"*Das ist die Widmung / This is the dedication (optional)*"

Acknowledgement

Das ist die Danksagung / This is the acknowledgement (optional)

Abstract

Das ist die Kurzfassung (siehe Abschnitt **??**) / This is the abstact (see section **??**).

Contents

Nomenclature

Greek symbols

Abbreviations and Acronyms

BVP	Blood Volume Pressure
Cobots	Collaborative Robots
DA-DCW	Dynamic Collision Avoidance with Direct Collaboration in a Shared Workspace
DA-SHW	Dynamic Collision Avoidance in a Shared Workspace
EDA	Electrodermal Activity
GMM	Gaussian Mixture Model
GSR	Galvanic Skin Response
HR	Heart Rate
HRV	Heart Rate Variability
KNN	K-Nearest Neighbors
MPC	Model Predictive Control
NA-DCW	No Collision Avoidance with Direct Collaboration in a Shared Workspace
NA-SHW	No Collision Avoidance in a Shared Workspace
NA-SW	No Collision Avoidance in a Separated Workspace
NASA-TLX	National Aeronautics Space Administration–Task Load Index
PA-DCW	Predictive Collision Avoidance with Direct Collaboration in a Shared Workspace
PA-SHW	Predictive Collision Avoidance in a Shared Workspace
PPG	Photoplethysmography
PSS	Perceived Stress Scale
SAM	Self-Assessment Manikin
SVM	Support Vector Machines
TSST	Trier Social Stress Test

Usage of generative AI models

$^{\ddagger_{MO}}$	Media optimization: Correction, optimization, or restructuring of entire passages
$^{\ddagger_{MG}}$	Media generation: Creating entire passages from given content.

Explanations for the usage of generative AI models and its notation:
The bottommost level at which the identification is presented regarding the possible uses of generative AI models are subchapters of the 2nd order (e.g., 1.1.1, which may also appear without numbering), as otherwise, the identification would disrupt the reading flow due to frequent occurrences. Algorithms used for implementing generative AI models

are mentioned at least in the text or provided as pseudo-code to facilitate appropriate identification.

1 Introduction

1.1 Motivation ‡MO

Industry 4.0, also known as the Fourth Industrial Revolution, has brought about significant transformations, particularly in the manufacturing sector. This revolution has been characterized by the introduction of intelligent technologies such as the Internet of Things (IoT), cloud connectivity, big data, and human-robot collaboration. These advancements have led to notable improvements and innovations, with the core principle and driving force of innovation in Industry 4.0 being the enhancement of efficiency and productivity. Human-robot collaboration, a key component of Industry 4.0, has played a massive role in this advancement by bringing humans closer together and facilitating more efficient and cooperative workflows.

Traditionally, industrial robots like robot manipulators, autonomous mobile robots, and gantry models have been kept separate from human workers primarily due to safety concerns. These robots, with their large size, substantial weight, and high speed, pose potential hazards when in close proximity to humans. This traditional approach prioritized the physical separation of robots and humans in industrial settings. However, advancements in Industry 4.0 have significantly increased the use of collaborative robots (cobots), bringing them closer together to accomplish tasks jointly. This evolutionary progression has witnessed the transformation of robots from being secluded behind safety barriers to now operating side-by-side with their human counterparts, effectively capitalizing on their unique capabilities which combine human adaptability and decision-making skills with the precision and consistency offered by robots.

Looking towards the future of the emerging Industry 5.0, the focus shifts towards a more human-centric approach (Pereira and Santos 2023). Industry 5.0 aims to strike a balance between technological advancements and human needs and interests. The goal is to merge the technological efficiency of Industry 4.0 with a greater emphasis on enhancing human operator's well-being and satisfaction. Industry 5.0 seeks to address the human challenges related to Industry 4.0 by prioritizing the well-being of workers and placing them at the center of the manufacturing process.(Nahavandi 2019). This shows a significant shift from purely efficiency-driven operations to those that also prioritize human factors and environmental sustainability.

With a focus on the human centrism aspect of Industry 5.0, this thesis aims to delve into

the human aspect of human-robot interaction, considering how proximity to robots might affect the operator's physiological state. It aims to investigate how continuous interaction with robots impacts the stress levels experienced by humans and emphasizes the importance of monitoring and accurately assessing stress levels in human-robot collaborative environments. Sauppé and Mutlu (2015) have previously indicated that cobots have the ability to influence the mental states of human workers, as they are often perceived as social entities. The close proximity of humans to robots in the workplace can lead to heightened stress levels, mainly if the robot's movements appear to be potentially harmful (Lasota and Shah 2015). For instance, if a co-robot swiftly moves towards a worker or follows an unpredictable path, it may induce feelings of anxiety or fear due to the perceived risk of sustaining an injury. This, in turn, can negatively impact both productivity and the efficacy of human-robot collaboration. Furthermore, it can impede the complete utilization of the advanced capabilities offered by collaborative robots. As robots become more autonomous and capable, identifying and addressing these stress factors is critical to optimizing human-robot collaboration for enhanced productivity and making the working environment effective, efficient and safe.

1.2 Aim of the Thesis‡MO

The primary objective of this thesis is to evaluate the impact of varying collision avoidance strategies on human stress levels within the context of human-robot interaction. This involves conducting a study to collect and analyze data, aiming to understand the varying stress levels concerning different robot collision avoidance strategies while considering various collaboration levels and robot control strategies. Subsequently, the data obtained is utilized to develop a predictive model capable of identifying and addressing sources of stress during human-robot collaboration.
Specifically, the primary objectives of the thesis are:

- **Assessment of Human Stress Levels** : Develop a holistic approach for evaluating stress levels in human-robot interactions during different collaboration levels and robot control strategies, combining both objective physiological measures and subjective experiences. Objectively, the study will employ various physiological indicators such as Galvanic Skin Response (GSR), Electrodermal Activity (EDA), Heart Rate (HR), and body posture analysis. These indicators will provide quantifiable data on the body's physiological response to robot interactions. Subjectively, the study aims to incorporate personal feedback from participants gathered through questionnaires. This will offer insights into their personal feelings and perceptions regarding their interactions with robots. By blending these objective and subjective methods, the study aims to understand stress in human-robot interactions comprehensively. This entails designing an acquisition system that successfully takes data from several sensors at different frequencies and synchronizes it. Devices such as the Empatica E4 wristband are used for gathering data on GSR, EDA, and other parameters, as well as a motion capture system to record human posture and movement. A vital aspect would be to synchronize these many data streams, ensuring accurate and consistent assessment of human physiological states across different robot interaction

scenarios. An important aspect was designing and conducting a subject study to collect data on participants' physiological responses while doing different assembly tasks under different human-robot interaction scenarios. These scenarios included three distinct levels of robot collision avoidance strategies: No collision avoidance, dynamic collision avoidance, and predictive collision avoidance, as well as three different collaboration levels: separated workspace with the cobot, shared workspace, and shared workspace with direct collaboration. The aim is to gather comprehensive data to analyze the impact of these varying robot control strategies on human stress levels.

- **Stress Prediction Model**: Developing a model for predicting and classifying stress levels during human-robot collaboration. This model trained on the dataset of human physiological responses collected from the subject study. Various preprocessing and feature engineering techniques are used to prepare the data for the model. Various machine learning models, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and others, are evaluated to determine the best model for predicting stress levels. The ultimate goal of this model is to provide a reliable tool for real-time monitoring and assessment of stress levels during human-robot collaboration. Accurately identifying stress-inducing factors and patterns can contribute to the development of proactive interventions and collision avoidance strategies that can enhance the overall well-being and performance of individuals engaged in human-robot interaction scenarios.

1.3 Related Work

This and this and this have all done a comprehensive review of various stress detection and assessment methods. This literature review from our part relies on the knowledge gained by these authors.

1.4 Structure of the Thesis

In Chapter 2, we delve into the fundamental concepts of stress and examine the relationship between stress and key physiological signals such as EDA, HR and GSR. Chapter 3 describes the data collection process of the subject study that was conducted to collect data on participants' physiological responses. Chapter 4 describes the data analysis process of the subject study to analyze the collected data and the different pre-processing techniques and feature selection used to build the model. Chapter 5 describes the results of the subject study and the different machine learning models that were evaluated to determine the best model for predicting stress levels. Chapter 6 concludes the thesis and outlines the future work and research directions.

2

Theoretical foundation

2.1 Stress Definition and Measurement[‡MO]

Stress, a term frequently used in everyday language aswell as in scientific domain, is an individual's response to situations perceived as challenging, threatening, or overwhelming. Stress is an unpleasant emotional state that individuals experience when confronted with demands that they perceive as taxing or exceeding their coping capabilities (Lee et al. 2004).

Stress, also called the "fight-or-flight" response, is an evolutionary adaptation that equips someone to respond to demanding situations rapidly. When faced with a potential threat or challenge, the human body instinctively readies itself for self-defence or swift evasion. The body's sympathetic nervous system is responsible for this response, which rapidly increases the production of stress hormones like cortisol, adrenaline, and noradrenaline (Gedam and Paul 2021).

The hormonal changes cause a range of bodily reactions, including acceleration of the heartbeat, muscle tension, changes in posture, increased blood pressure, rapid breathing, and heightened sensory alertness etc. One can objectively measure these bodily changes, which generally fall into two categories: physical and physiological changes.

Physical measures focus on observable bodily changes that occur under stress. These include alterations in facial expressions, variations in the rate of eye blinking and pupil dilation, changes in body posture and movement patterns. These visible markers offer insights into an individual's stress levels.

Physiological measures, in contrast, involve using sensors to detect internal bodily changes indicative of stress. A range of biomarkers is employed for this purpose, including Heart Rate Variability (HRV) and HR, GSR electrodermal activity, respiratory patterns and cortisol levels. These biomarkers provide a more direct and quantifiable insight into the body's response to stress, making them valuable tools in stress assessment.

Experts specializing in research also meticulously design standardized questionnaires for the subjective evaluation of stress, which has been a longstanding approach to understanding individual stress levels. These questionnaires are structured to accurately capture an individual's perceived stress levels and their reactions to various stressors. This subjective methods are crucial as they offer insights into the personal experiences and perceptions of stress, which may not always be evident through objective measures.

2.2 Subjective Measures ‡MO

Subjective ratings, such as self-report questionnaires, have been commonly used as a direct method to estimate levels of mental stress in humans in an experimental setting. (Aigrain et al. 2018). Participants are asked to answer a variety of questions about their experiences in the experiment. There have been a different variety of questionnaires and tests used to investigate the emotional state or perceived stress from the human participants in experimental setting. Some of the most widely used ones are Self-Assessment Manikin (SAM) (Bradley and Lang 1994), National Aeronautics Space Administration–Task Load Index (NASA-TLX) (Hart and Staveland 1988), Perceived Stress Scale (PSS) (Cohen, Kamarck, and Mermelstein 1983) etc.

The SAM is a non-verbal pictorial assessment technique designed to measure emotional response and affective reaction to diverse stimuli (Bradley and Lang 1994). Administered typically at the conclusion of each experimental task, it asks participants to assess their emotions and affective state on a scale from 1 to 9 across three dimensions: valence (the nature of the emotion, ranging from positive like relaxation to negative such as fear), arousal (the intensity of the emotion), and dominance (the extent to which the emotion is perceived as controllable).

NASA-TLX is extensively used in various research studies to evaluate mental stress levels. Notably, Nguyen and Zeng (2017) implemented this tool to assess the mental workload of surgeons during endoscopy training. In a similar vein, Zakeri et al. (2023) applied NASA-TLX within the context of smart factories to scrutinize factors like task complexity, time pressure, and collaboration duration, all contributing to mental stress.

Primarily, NASA-TLX aims to measure perceived workload across different tasks, particularly in high-stress environments. It seeks to capture a comprehensive picture of stress and workload through subjective user experiences. The tool does this by evaluating six key dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration, each playing a role in the overall task load assessment for an individual. The evaluation process incorporates a NASA-developed technique for gauging the relative significance of these factors in the experienced workload. For each of the six dimensions, NASA-TLX employs a 21-point bipolar scale, allowing participants to range their workload assessment between two extremes, such as "Low/High." Furthermore, the process involves presenting pairs of rating scale labels, prompting subjects to choose which label is more pertinent to their cognitive workload experience in the task. This selection pattern enables the assignment of a weight to each cognitive load factor, culminating in an overall score that aligns with a specific subject's experience.

Nguyen and Zeng (2017) demonstrates the use of the NASA-TLX for assessing mental stress by interpreting workload measurements as indicators of stress levels. This approach highlights NASA-TLX's utility in evaluating how task demands can translate into mental stress.

In our research, NASA-TLX emerged as the most fitting tool for assessing stress in the context of the external stimuli of human-robot collaborative tasks. Stress, as a multifaceted experience, varies subjectively in perception. It might arise in response to workload (our focus area) but also encompasses factors such as emotional responses, individual coping mechanisms, and various personal and environmental influences. Characterized often by

feelings of strain, anxiety, or pressure, stress responds to both internal and external stimuli. Given our specific objective to assess stress related to the external stimuli of the task, other questionnaires like the Perceived Stress Scale (PSS), Trier Inventory for the Assessment of Chronic Stress (TICS) (Schulz and Schlotz 1999) - which measures general chronic stress over a period, or the Daily Stress Inventory (DSI) (Brantley et al. 1987) - evaluating daily stress events, did not align with our requirements.
While subjective questionnaire is a powerful tool to measure stress levels directly, it is important to mention their limitations. The reliance on self-reporting means it's subject to individual biases and may not accurately reflect real-time stress levels. People might not always be able to accurately introspect and report their feelings or may inadvertently skew their responses based on what they think researchers want to hear.

2.3 Objective Measures ‡MO

Objective measures of stress include measuring of physiological and physical measures by means of sensors. These sensors are either placed on the human body to measure bodily changes in an unobtrusive manner or at a distance in case of physical measurements. Objective measures of stress are free from human intervention and hence cannot be biased. The two sensors we used to collect biosignals objectively are the Empatica E4 and the OptiTrack Motion Capture System. They are introduced below:

2.3.1 Empatica E4

The Empatica E4 (Empatica n.d.[a]) (see Figure 2.1) wristband is a versatile and compact device designed to capture a wide range of physiological data in real time. It has four sensors: a Photoplethysmography (PPG) sensor, which measures Blood Volume Pressure (BVP); an EDA sensor, which is used for measuring GSR; a 3-axis Accelerometer to capture motion-based activity and an infrared thermopile to reads skin temperature (ST) (Garbarino et al. 2015). Its unobtrusive nature makes it comfortable to wear, while its comprehensive data collection capabilities have made it an invaluable asset.
The Empatica E4 wristband collects BVP data using the PPG with a process that involves emitting green and red light from LEDs into the skin and measuring the reflected light with a sensor(see Figure 2.2a). The green light, absorbed by the blood, provides a pulsatile signal corresponding to the cardiovascular pulse wave used to determine heartbeats. The red light acts as a reference to correct for motion artifacts. Algorithms then process this data within the wristband to output the BVP, from which the interbeat interval (IBI)—the time between heartbeats—is calculated, offering a non-invasive method to monitor heart rate continuously.(Empatica n.d.[c])
EDA is measured by detecting the electrical conductance across the skin, which is an indirect indicator of the sweat gland activity influenced by the sympathetic nervous system. To obtain these measurements, Empatica employs a method that relies on passing a minimal electrical current between two electrodes that are in contact with the skin, typically placed on the bottom wrist.
The wristband also includes a 3-axis accelerometer and an infrared thermopile, which can track body temperature and movement, providing a comprehensive overview of the

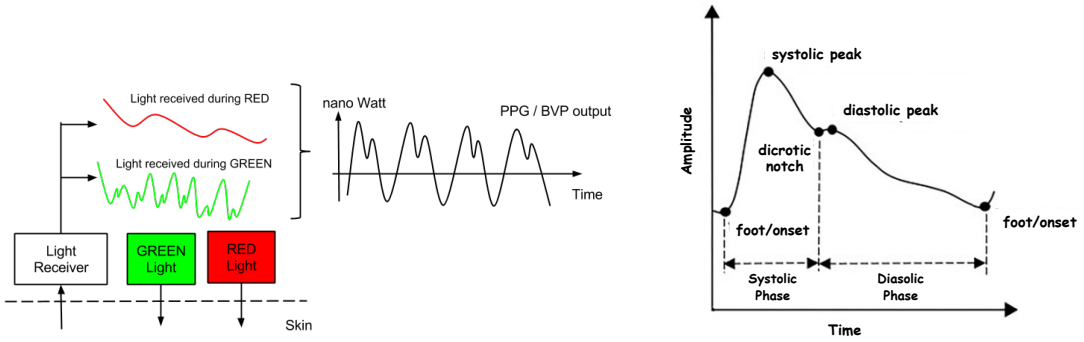
wearer's physiological state.

Figure 2.1: Empatica E4 features (Empatica n.d.[b])

Photoplethysmogram-PPG

Photoplethysmogram (PPG) also known as Blood Volume Pulse (BVP) are non-invasive optical techniques used to monitor changes in blood volume. They rely on the principles of light absorption and reflection to capture valuable information about cardiovascular activity. PPG sensors commonly found in wearable devices obtain BVP signals by transmitting light into the skin and measuring the amount of light either transmitted through or reflected back.(V. Roberts 1982)
When the heart beats, it propels blood through the circulatory system, causing periodic changes in the volume of blood vessels. PPG sensors emit light into the tissue and measure the amount of light that is either absorbed or reflected back. During each heartbeat, blood absorbs more light, leading to a decrease in the amount of light detected by the sensor. Between heartbeats, when blood flow is less pulsatile, more light is detected.(Zhang et al. 2001)
The resulting waveforms from PPG typically consist of a series of peaks and troughs, with each peak corresponding to a heartbeat (systole) and each trough representing the resting period between beats (diastole). By analyzing the time intervals between these peaks, the heart rate can be calculated. This heart rate measurement is fundamental and provides valuable information about a person's cardiovascular health and overall fitness level. It serves as a key metric in various applications, including exercise tracking, medical diagnosis and in our case here stress assessment.
Furthermore, PPG signals enable the assessment of HR and HRV. HRV is the variation in time between successive heartbeats and is an essential indicator of the autonomic nervous system's activity. By analyzing the subtle changes in the intervals between PPG peaks, HRV can be quantified. High HRV typically indicates a healthy heart and a well-balanced autonomic nervous system, while reduced HRV can be associated with stress, illness, or various medical conditions. HRV analysis provides insights into the body's ability to adapt

to different situations and is valuable for assessing stress levels, mental well-being, and overall cardiovascular health.
Other measures that can be derived from PPG data include and estimation of blood oxygen saturation levels (SpO2), valuable for respiratory and circulatory health assessment. PPG can also be used to estimate respiration rate, reveal vasomotor activity changes associated with the autonomic nervous system, emotions, or vascular health, and provide insights into arterial stiffness and blood flow dynamics as well as blood pressure. First derivative and second derivates of PPG signals can also be analyzed. The first derivative (Velocity Plethysmogram, VPG) and the second derivative (Acceleration Plethysmogram, APG) features can be used for blood pressure estimation etc.(Suboh et al. 2022)



(a) PPG process (b) Typical PPG Waveform

Figure 2.2: (Empatica n.d.[b]) (Suboh et al. 2022)

Electrodermal Activity-EDA ‡MO

Electrodermal Activity (EDA), also known as the galvanic skin response(GSR), is a way to measure changes in how our skin conducts electricity. Even small quantities of sweating that are not visible on the skin might affect the electrical conductivity of the skin. As the body perspires, the skin's conductivity increases, allowing for the measurement and inference of physiological or psychological states. EDA particularly measures the variations in the skin's electrical conductivity, which are influenced by the amount of sweat produced by the eccrine sweat glands. The secretion of sweat is primarily induced by the activity of the central nervous system, which is influenced by emotional and cognitive states(Machado-Moreira et al. 2008). Thus, EDA becomes one of the promising noninvasive methods widely used in detecting stress and emotion. EDA is a powerful method for real-time measurement and could be used as an index of emotional or cognitive stimulation related to stress.(Gellman 2020).EDA is useful in several ways: it shows how we respond emotionally, helps us see how our body reacts to stress etc. It acts as a biomarker for emotional responsiveness and serves as a key indicator for stress-related bodily responses.
EDA consists of two primary components: the tonic component and the phasic component. The tonic component, sometimes referred to as skin conductance level (SCL), represents gradual and constant changes in the background of the signal. On the other hand, the phasic components, known as skin conductance response (SCR) or spontaneous fluctuation of skin response, are the swift and brief changes within the signal that happen at certain

time intervals (2017). SCR appears in response to stimuli activating the sympathetic nervous system. Consequently, SCR can be linked to a stimulus and can be valuable in measuring cognitive stress levels. However, directly extracting the components of EDA isn't straightforward.

When EDA sensors measure skin conductivity (SC) signals, they typically yield results in microsiemens. To extract the SCL and SCR components accurately, it is necessary to deconvolve the SC signals (Alexander et al. 2005, postnote). Without proper separation of the original SC signals, overlapping SCRs can lead to less precise information during feature extraction . Therefore, it is crucial to perform deconvolution to distinguish the SCR and SCL signals effectively.

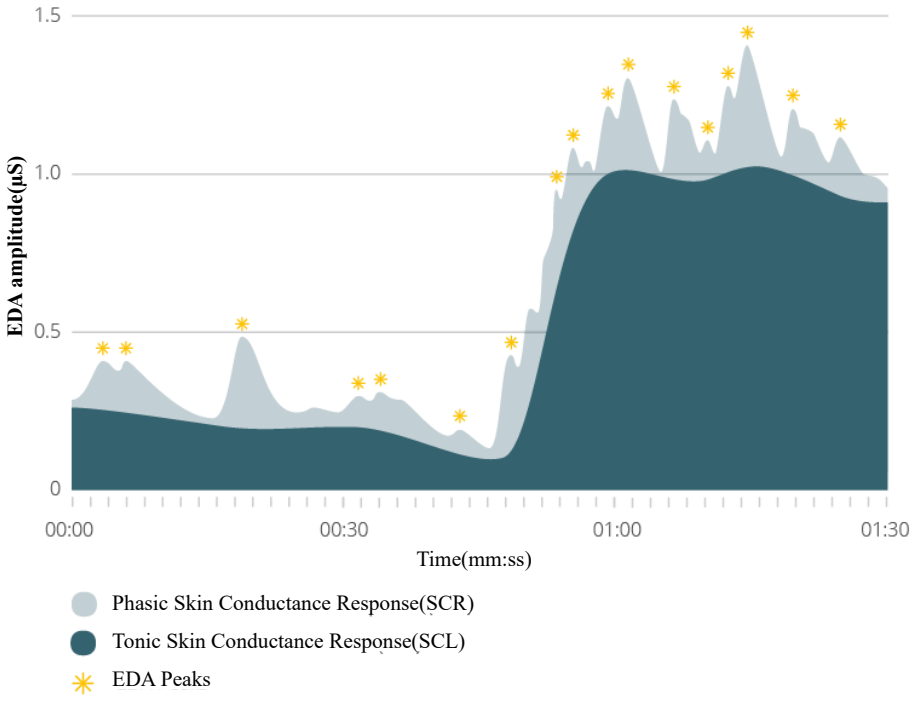


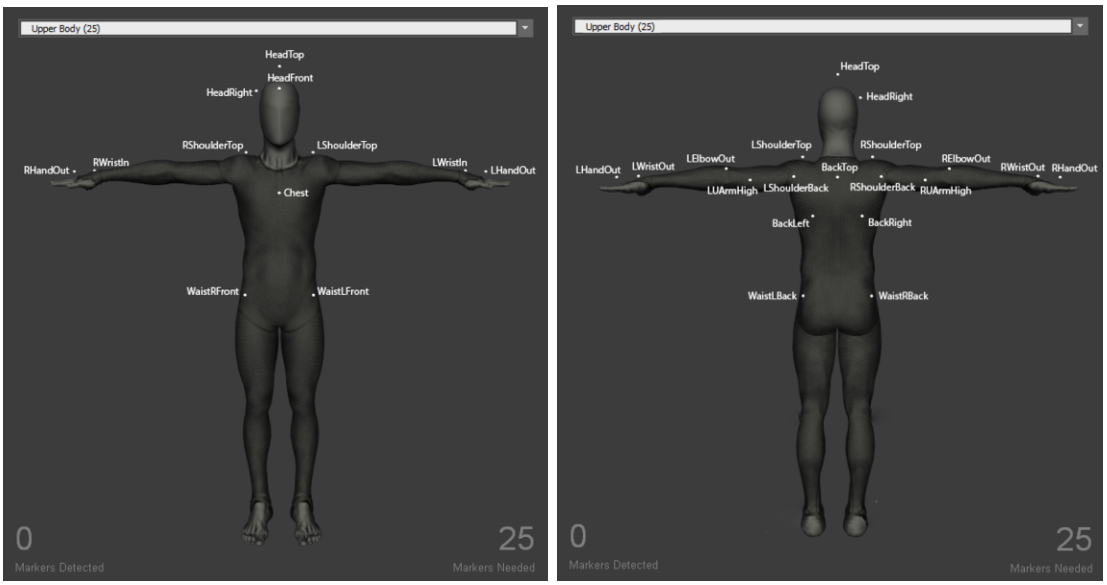
Figure 2.3: EDA Example signal - (imotions n.d.)

Skin Temperature

Other Physiological signal measured by the Empatica E4 is the skin temperature. Typically, the normal range for skin temperature lies between 33.5 and 36.9°C but changes in skin temperature can be connected to the stressful and anxious conditions (Empatica n.d.[c]) Skin temperature is strongly correlated to the heart activity and sweat reaction of an individual, so when a person is sweating, their skin temperature increases thus being a measure of increased stress levels.

2.3.2 Motion Capture‡MO

The OptiTrack Motion Capture System is a motion capture system that uses an array of 12 high-speed cameras equipped with advanced optics and infrared sensors. These cameras are positioned to cover a designated area, creating a three-dimensional space where every movement is tracked and recorded. The system detects reflective markers placed on key points of a subject's body. As the subject moves within the camera's field of view, the system tracks the spatial position and orientation of these markers, seamlessly translating physical movements into digital data.
For tracking the human body, the system uses a set of 25 marker points, which are placed at strategic points as shown in Figure 2.4. This configuration ensures a thorough capture of the upper body movements. The calibration process is a critical step where each marker on the subject's body is meticulously mapped onto a digital skeleton model. This mapping ensures that the system can accurately track the movements of each marker in relation to the body's overall structure.The system individually tracks each marker as the subject moves, allowing for a detailed representation of motion. The Motive software, integral to the OptiTrack system, plays a key role here. It enables users not only to visualize the movements in real time but also to record and analyze the data. The software translates the positional data of the markers into a skeletal animation, offering a clear and dynamic representation of the subject's movements.



(a) Front View (b) Back View

Figure 2.4: 25 Upper Body Marker Set (OptiTrack n.d.)

2.4 UR10 robot and Collision Avoidance ‡MO

The UR10 is part of the Universal Robots family of Collaborative Robots (Cobots), designed to collaborate directly with humans in a shared workspace. This robotic arm with six joints is highly flexible, has a reach of 1300 mm, and can handle a payload of up to

10 kg, making it suitable for a wide range of applications and collaborative tasks.Figure 2.8 shows the ur10 with its 6 rotatory joints.
Despite its capability to perform complex tasks, the UR10 does not inherently possess collision avoidance strategies. Its default response to encountering a collision is typically to halt operations to prevent any damage or injury, relying on built-in safety features that comply with industrial safety standards. However, innovative research and development have sought to enhance the UR10's capabilities with advanced collision avoidance strategies.
It is necessary to simplify the human body tracked using the motion capture system in **??** for ease of computation in collision avoidance trajectory planning. The simplified human body is modelled using line swept spheres (Larsen et al. 2000). A swept sphere, also known as a capsule, is the volume of a sphere that is swept along a straight line. It resembles a cylinder in shape, but it is simpler for distance metric computational purposes.

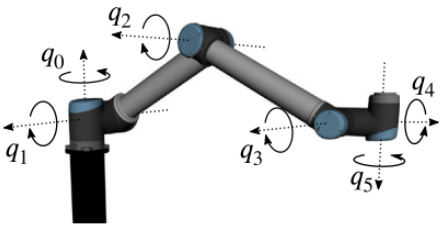


Figure 2.5: 6 DOF UR10 taken from (Krämer et al. 2020)

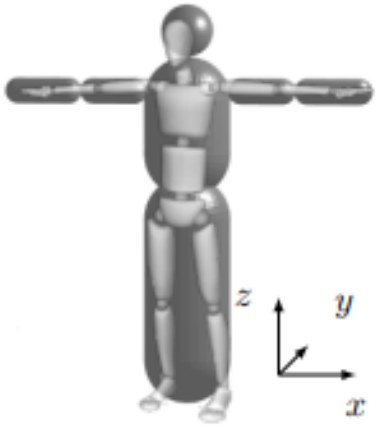


Figure 2.6: Simplified human model taken from (Renz, Krämer, and Bertram 2023a)

Seven spherical volumes make up the human body model: a sphere represents the head, while these swept sphere volumes represent the rest of the body parts, such as the arms, the torso and the parts below the hip is depicted by these swept spheres. This method effectively and precisely captures the human form and its movements in a simplistic manner, reducing complexity and increasing computational speeds. Figure 2.9 shows the simplfied human body model superimposed on the skeleton model of the motion capture system. The motion capture system continuously monitors and records the individual's movements. This information is then fed into the robot's planning system, allowing it to maneuver and modify its actions adeptly in response to the human's changing position and movements.
This thesis discusses three different levels of collision strategies.
The first one already discussed is the default collision strategy of the UR10 robot, where it stops at a collision.
The second one uses dynamic collision avoidance, which tracks moving humans and tries

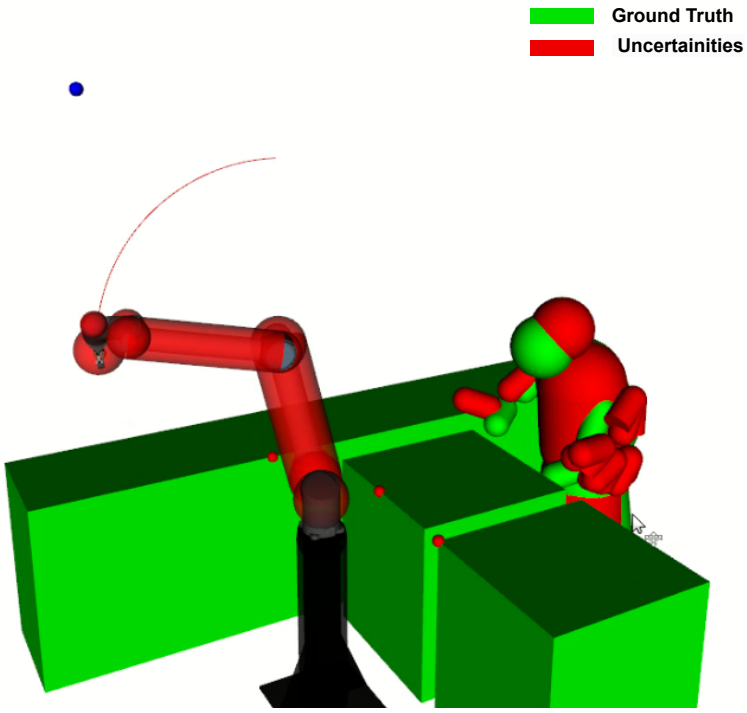


Figure 2.7: Predictive Collision Avoidance

to avoid them in real-time. It focuses Model Predictive Control (MPC), a method that anticipates and adapts to changes in the environment, including moving obstacles. This control scheme by Krämer et al. (2020) integrates online trajectory optimization with MPC. This approach enables the robot to adjust its movement in real-time, considering potential collisions and task variations. The prediction model within MPC approximates the robot's joint velocities and positions, facilitating efficient adaptation to dynamic obstacles. The control system is structured in a cascaded manner. The outer loop handles MPC, planning optimal movement trajectories, while the inner loop consists of tracking controllers for velocity references. This architecture effectively decouples the complex dynamics of robot motion control, simplifying the computational process.(Krämer et al. 2020)

The third predictive collision avoidance strategy is predicting the future human motion. One approach for predicting human motion is the extrapolation of the human skeleton's joint states in joint space using Polynomial Estimation (PE) methods as done by Renz, Krämer, and Bertram (2023a). This technique involves fitting a polynomial to past joint angles in a least-squares sense and predicting future joint states, which includes the joint angles and their velocities and accelerations. These predictions inherently come with some degree of uncertainty, especially as the prediction horizon extends further into the future. To represent this uncertainty, a Gaussian Mixture Model (GMM) is applied, which describes multiple potential future extrapolations. Each extrapolation is defined by observations of the errors between the predicted joint states and the actual (ground truth) joint states at different time steps. These errors are time-dependent, with the assumption that errors increase for predictions further into the future.(Renz, Krämer, and Bertram

2023b).
The GMM consists of several components, each representing a normal distribution with its own mean and covariance matrix. These components collectively form a probabilistic model of the potential errors in joint state predictions over time. The parameters of the GMM are updated using an Expectation-Maximization (EM) algorithm, which maximizes the likelihood of the observed data (the past prediction errors). However, due to computational constraints, the GMM is not updated with every new extrapolation but rather at set intervals considering the most recent set of extrapolations.
This GMM approach allows for the estimation of uncertainties in a real-time capable manner, which is crucial for adjusting the robot's motion plan to avoid collisions with humans dynamically and safely.

2.5 Stress Classification ‡MG

The classification of stress levels in individuals has been the subject of extensive research, leading to the development of various methods. Stress classification is a supervised learning problem, a category of machine learning. In supervised learning, the model is trained on a labelled dataset, meaning each input data point is associated with a known output label. In the context of stress classification, these labels represent different stress states, such as 'stressed' or 'not stressed'. The model learns from this training data, enabling it to make predictions or classify new, unseen data based on recognised patterns.
Among the various techniques employed for stress classification, Support Vector Machines (SVM), K-nearest neighbors (KNN), Logistic Regression, and Decision Trees are notable for their widespread use. These methods each have unique foundational concepts and operational mechanisms:

- **Support Vector Machines (SVM)**: SVMs are known for their efficacy in classifying non-linearly separable data. They function by identifying the optimal hyperplane that separates different classes in the feature space.

- **K-Nearest Neighbors (KNN)**: KNN classifies data based on the similarity principle, considering how closely a new data point resembles existing points in the training set. This method is particularly useful when dealing with irregular decision boundaries.

- **Naive Bayes Method**: This method applies Bayes' Theorem with the assumption of independence among predictors. Naive Bayes classifiers are fast and efficient, especially useful for large datasets and scenarios where the features are conditionally independent.

- **Random Forest**: As an ensemble method, Random Forest combines multiple decision trees to improve the model's accuracy and robustness. It is highly versatile and can handle both classification and regression tasks effectively, especially beneficial in dealing with overfitting issues common in individual decision trees.

- **Neural Network (MLP)**:Multilayer Perceptrons, a type of neural network, consist of multiple layers of interconnected nodes or neurons, where each layer's output is the input for the next layer. MLPs are particularly powerful in capturing complex patterns in data, making them suitable for a wide range of classification tasks, including those involving high-dimensional data.

Research by Bhushan and Maji (2023), Gedam and Paul (2021), Vos et al. (2023), and Sharma and Gedeon (2012) has been instrumental in reviewing, summarizing, and comparing these commonly utilized methods in stress classification. Their analyses provide insights into the efficiency, applicability, and specificities of these classifiers, offering a deeper understanding of their role in stress level classification. This section aims to present an overview of the background information of these chosen classifiers.

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification and regression tasks. At its core, SVM aims to find the optimal hyperplane that separates data points of different classes in a high-dimensional space. This separation is achieved to maximise the margin between the data points of other classes, ensuring the best possible classification.
The process of selecting the best hyperplane for data separation follows a methodical approach, adaptable for both binary and multiclass scenarios. Here's a general overview of the steps involved:
To separate data points into distinct classes, SVM employs kernel functions, which can be denoted as $K(\mathbf{x}_i, \mathbf{x}_j)$. These kernel functions transform the input data into a higher-dimensional space where a hyperplane can be used for separation.
The support vectors are the data points closest to the hyperplane and satisfy the following conditions:

$$\mathbf{w} \cdot \mathbf{x}_+ + b = +1, \quad \text{for positively labeled data,} \tag{2.5.1}$$

$$\mathbf{w} \cdot \mathbf{x}_- + b = -1, \quad \text{for negatively labeled data.} \tag{2.5.2}$$

These points are vectors $\mathbf{x}_+$ and $\mathbf{x}_-$ that lie on the boundary of the margin.
The margin is defined as the distance between the support vectors and the hyperplane. It can be calculated as:

$$\text{margin} = \frac{2}{\|\mathbf{w}\|}. \tag{2.5.3}$$

The optimal hyperplane is the one that maximizes the margin. The objective function to maximize the margin while classifying the training data correctly is given by:

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}, \quad \text{s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \forall i. \tag{2.5.4}$$

This can also be equivalently written as a minimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \quad \text{s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \forall i. \tag{2.5.5}$$

By solving this optimization problem, SVM finds the optimal hyperplane that classifies the data with the maximum margin, enhancing the generalization ability of the classifier. The SVM stands out from other classifiers that also use lines or hyperplanes due to its strategy of utilizing the maximum margin separating hyperplanes. By focusing on this maximum margin, the SVM enhances its ability to correctly predict the classification of new, previously unseen instances. The chosen hyperplane effectively determines how an unknown sample is classified, falling into one class or another based on which side of the hyperplane it lies. This approach ensures that the classifier is not only effective but also robust in its predictions, making SVM a valuable tool in a wide range of classification applications.
Support Vector Machines (SVMs) are particularly well-suited for the task of stress classification from physiological data. By constructing an optimal hyperplane in a high-dimensional feature space, SVMs can efficiently differentiate between stressed and non-stressed states based on various biosignals. The mathematical foundation of SVM in the context of stress classification can be represented as follows:
Below is the table of common kernel functions used in SVM:

Table 2.1: Kernels

Kernel	**Expression**
Linear Kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$
Polynomial Kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + c)^d$
Sigmoid Kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta \mathbf{x}_i^\top \mathbf{x}_j + \theta)$
RBF	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

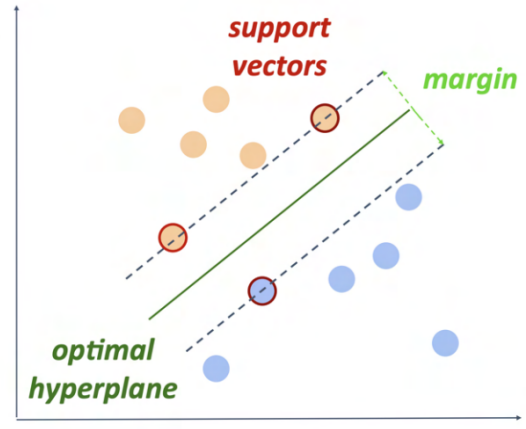


Figure 2.8: SVM taken from (Badillo et al. 2020)

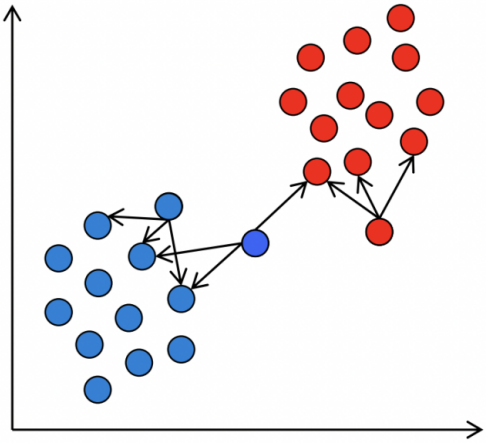


Figure 2.9: kNN taken from (Badillo et al. 2020)

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a versatile algorithm used in supervised machine learning, predominantly employed for classification and, to some extent, regression tasks. It operates

on the principle of classifying new data points based on the majority class among its nearest neighbors in the feature space. The proximity of neighbors is determined using distance metrics, such as the Euclidean distance and Manhattan distance, calculated by the following formulas:

$$\text{Euclidean Distance}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{2.5.6}$$

$$\text{Manhattan Distance}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}| \tag{2.5.7}$$

The choice of 'k', the number of nearest neighbors to consider, is critical. A smaller 'k' can make the model sensitive to noise, while a larger 'k' can lead to high computational costs and may include less relevant neighbors. Optimal 'k' is often determined through cross-validation.
Some variations of KNN employ weighted voting, where closer neighbors have more influence on the classification than more distant ones, enhancing the accuracy in certain scenarios.
KNN is effective for stress classification in datasets where stress indicators form distinct clusters. In physiological data, for instance, patterns of stress responses might cluster, enabling KNN to differentiate between stressed and non-stressed states. While KNN's simplicity and effectiveness are advantageous for smaller datasets, it faces challenges like high computational cost in large datasets and sensitivity to irrelevant features. Feature selection and dimensionality reduction techniques are employed to mitigate these issues.
In conclusion, KNN's approach of classifying data based on nearest neighbors, along with its adaptability to distance metrics like Euclidean and Manhattan distances, makes it a valuable tool for stress classification. Key considerations for its effective application include dataset size, feature relevance, and the optimal choice of 'k'.

Naive Bayes Method

Naive Bayes is a highly efficient and straightforward algorithm in supervised machine learning, predominantly used for classification tasks. The algorithm is based on Bayes' Theorem and assumes independence among predictors. Despite its simplicity, Naive Bayes can be remarkably effective, hence the name "Naive" Bayes.
The algorithm applies Bayes' Theorem, which provides a way to calculate the probability of a hypothesis based on prior knowledge. The theorem is formulated as:

$$P(Y|X_1, ..., X_n) = \frac{P(Y)P(X_1, ..., X_n|Y)}{P(X_1, ..., X_n)}$$

where Y is the class variable, and $X_1, ..., X_n$ are the dependent feature variables.
The key "naive" assumption in this classifier is the conditional independence of features:

$$P(Y|X_1, ..., X_n) = \frac{1}{Z}P(Y)\prod_{i=1}^{n} P(X_i|Y)$$

Here, Z is a scaling factor, $P(Y)$ is the prior probability of class Y, and $P(X_i|Y)$ is the likelihood of feature i given class Y.
There are different types of Naive Bayes classifiers, including Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, each suited for different types of feature distributions.
Naive Bayes is particularly known for its simplicity, speed, and efficiency in handling large datasets. It performs well in scenarios with categorical input variables compared to numerical variables. For stress classification, Naive Bayes can be effective, especially when the dataset features are conditionally independent.
Despite its simplicity, Naive Bayes can outperform more complex models, particularly when the assumption of independence holds. However, this assumption is also the biggest limitation, as it rarely holds in real-world data. It also struggles with zero-frequency problems, where if a categorical variable has a category in the test data set that was not observed in the training data set, the model will assign a 0 probability and will be unable to make a prediction.
In conclusion, the Naive Bayes method, with its foundation in probability theory and its operational efficiency, is a robust tool for classification tasks. Its performance is particularly notable in stress classification when the dataset features adhere to the independence assumption. Proper understanding and handling of its assumptions and limitations are crucial for achieving optimal performance.

Random Forest

Random Forest is an ensemble learning method widely used in supervised machine learning for both classification and regression tasks, though it is particularly known for its effectiveness in classification, including multi-class scenarios. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
A key feature of Random Forest is its use of bagging (bootstrap aggregating) and feature randomness when building each individual tree to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. This process involves creating multiple subsets of the original dataset (with replacement), training a decision tree on each, and then averaging the results to improve the model's accuracy and robustness.
One of the primary advantages of Random Forest is its ability to handle large datasets with higher dimensionality. It can manage thousands of input variables without variable deletion, making it highly effective for datasets with a large number of features. It also provides a good indicator of feature importance, which can be helpful in feature selection.
Random Forest is particularly effective in classification tasks, including multi-class classification. Its ability to handle complex interaction structures and non-linear relationships makes it suitable for diverse applications, including stress classification using physiological data. The ensemble nature of Random Forest helps overcome the overfitting problem often seen with individual decision trees.
Despite its advantages, Random Forest can be computationally intensive and may not perform well on datasets with a very large number of features relative to the number of

samples. However, its robustness, ease of use, and ability to run in parallel make it a popular choice for a wide range of classification tasks.
In conclusion, Random Forest's ensemble approach, combining multiple decision trees to produce more accurate and robust models, makes it a powerful tool in machine learning for classification tasks. Its application in multi-class classification scenarios, like stress detection using physiological signals, demonstrates its versatility and effectiveness in handling complex and high-dimensional data.

Neural Networks- Multilayer Perceptron

The Multilayer Perceptron (MLP) is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Except for the input nodes, each node is a neuron with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training, effectively enabling it to learn from data and improve its performance over time.
An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. This network architecture allows MLP to learn complex data patterns, making it highly effective for classification tasks, including multi-class classification. The presence of one or more non-linear hidden layers enables the MLP to learn non-linear functions, distinguishing it from a linear perceptron. It can distinguish data that is not linearly separable, which is a limitation of single-layer perceptrons.
MLP's are widely used in machine learning for their ability to analyze large datasets and their flexibility in fitting various data types. In the context of classification, including multi-class scenarios, MLPs can effectively capture complex relationships between input features and the target output. They are particularly useful in areas such as image recognition, speech recognition, and complex decision-making tasks, where the relationship between input and output is non-linear and intricate.
However, training MLPs can be a complex and computationally intensive task. The choice of the number of hidden layers and the number of neurons in each hidden layer is crucial and can significantly impact the model's performance. Overfitting is a common challenge in MLPs, often addressed through techniques like regularization and dropout.
In conclusion, MLPs, with their multi-layered neural structure and backpropagation training algorithm, are powerful tools in the realm of neural networks, suitable for a wide range of classification tasks, including those involving multi-class categorization. Their ability to model complex, non-linear relationships makes them invaluable in advanced machine learning applications.

3

Data Collection-Subject Study

A collaborative assembly task involving wooden pieces was designed in the robot laboratory of the Institute of Control Theory and Systems Engineering at the Technical University of Dortmund. This setup aims to accurately replicate tasks commonly seen in industrial settings where collaboration between humans and cobots is frequently observed. The Universal Robot UR10 was selected as the cobot to be used in the study. A call for participants invite was sent out, and 20 male students who all had a technical background volunteered to participate in the study. Seventeen of the students did not have any previous experience working with a robot in any way, whereas the remaining had some previous experience with robots. The mean age of the participants were 23 ± 2 years. Before the experiment began, all participants were given a comprehensive overview of the study's objectives and procedures, along with a consent form. Only those participants who agreed to the terms and signed the consent form were permitted to proceed with the experiment. To ensure the ethical integrity of the study, a prior request for ethical approval was submitted to the appropriate ethics council and permission was obtained to conduct the subject study.

3.1 Design of Tasks

For the purpose of studying the impact of stress on various factors, the assembly of different mock items using wooden children's toys was chosen as a method of replicating an industrial assembly task.
These included three distinct levels of different collaboration levels:

- **Seperated Worksapce**: Human and cobot have no overlapping workspace. The cobot works in the background.The human already has all items required for the assembly task in front of him.
- **Shared Workspace**: The human and cobot share the same work area.The cobot brings each item required for the assembly tasks to the human and places it on the table in front of the human.
- **Shared Workspace with Direct Collaboration**: The human and cobot share the same work area and the cobot brings the item required for the assembly tasks to the human and directly hands over the items to the human.

Collision Avoidance Strategy	Seperated Workspace (A)	Shared Workspace (B)	Shared Workspace with Direct Collaboration (C)
No Collision Avoidance (X)	AX	BX	CX
Dynamic Collision Avoidance (Y)	Not Required	BY	CY
Predictive Collision Avoidance (Z)	Not Required	BZ	CZ

Table 3.1: Task names for different Collision Avoidance Strategies and Workspace Scenarios

As well as three robot collision avoidance strategies:

- **No Collision Avoidance**: No collision avoidance measures are in place, and the robot stops at a collision.
- **Dynamic Collision Avoidance**: The robot identifies the human as a dynamic obstacle and adjusts its trajectory to avoid collisions.
- **Predictive Collision Avoidance**: This strategy uses predictions to predict the human's future position and adjusts its trajectory to avoid collisions.

The Table 3.1 shows the various combinations of factors and the naming conventions of the tasks, yielding seven experimental scenarios since collision avoidance is not applicable when Separated Workspaces are involved. For example, task BZ employed a predictive collision avoidance strategy while doing the task in a shared workspace. So a within-subject experimental design was employed where each participant was tested on all seven scenarios. By having the same participants perform each task, we minimized the impact of differing skill sets, experiences, and cognitive abilities, which could otherwise skew the results.

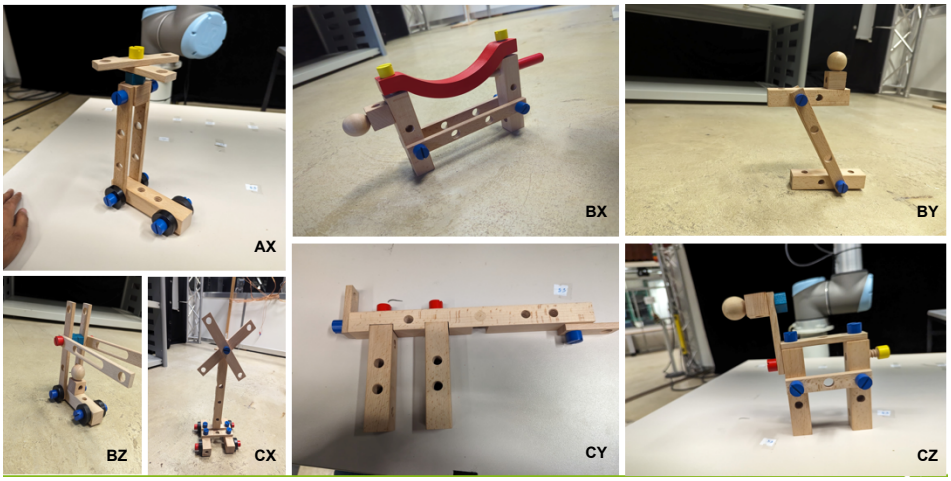


Figure 3.1: 7 different assembly tasks

To avoid any potential learning effects and ensure that each task measures the intended variables accurately, we had to design seven different assembly tasks. Each task involved assembling a unique item, chosen of similar complexity and required effort. This similarity in task difficulty was crucial to avoid the introduction of other variables, such as familiarity with the task, which could affect the results. Figure 3.1 shows the final assemblies completed in each of the seven tasks. Each task had approximately 4-6 different parts, which were supposed to be screwed together to complete the tasks. By standardizing the complexity across tasks, we aimed to isolate the impact of the collision avoidance strategies and collaboration levels. We also had to consider the order in which the task was administered for each participant, ensuring that learning or fatigue does not affect the outcomes in any way. Each participant experienced the tasks in a unique order, balancing out any potential biases introduced by the order of task presentation.

3.2 Apparatus and Experimental Setup

The experiment was set up in a specially designated area of our laboratory. Figure 3.2 shows the experimental setup. At the center of this arrangement was the collaborative workspace, featuring a table and chairs for the human participant (Area B in Figure 3.2), positioned directly opposite the robot's dedicated workspace. Adjacent to this, on the right side of the collaborative workspace, a table was placed to hold the various items needed for the assembly tasks(Area A Figure 3.2). The robot would pick the necessary items from this table for each specific task and deliver them to the human participant. In front of the human participant, a mobile device was also placed. This device was key to the experiment, as it presented the participant with concise, step-by-step instructions for each assembly task. Figure 4.4a shows how these instructions were visually displayed, offering clear and easy-to-follow guidance. The participant would start the assembly task as soon as the first item is delivered to the human participant. The robot would then proceed to the next item, and so on until all items are delivered to the human participant.

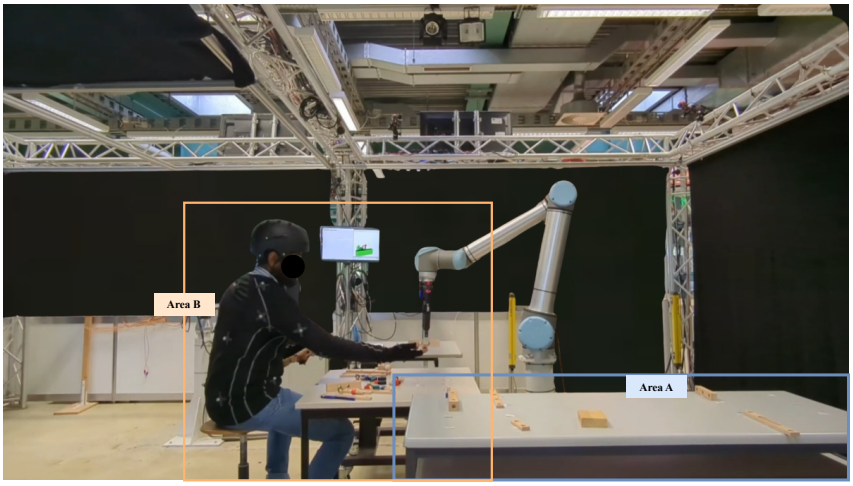


Figure 3.2: The experimental setup

The entire experimental area had an OptiTrack motion capture system, outfitted with 12 high-precision cameras fitted across all 4 sides as partly seen in Figure 3.3. These

OptiTrack cameras, known for their accuracy and low latency are used to capture every detail of the human participant's movements. This setup was necessary for the collision avoidance trajectory planning for the robot and also providing a detailed and continuous record of the participant's interactions with the robot. The use of the OptiTrack system enabled us to gather precise data on human motion. The participants were equipped with the motion capture suit which had 25 distinct marker points which were used to capture the human's head and upper body. To prioritize participant safety, especially given the proximity to a large robotic arm, a helmet also equipped with the head markers was provided to each participant.

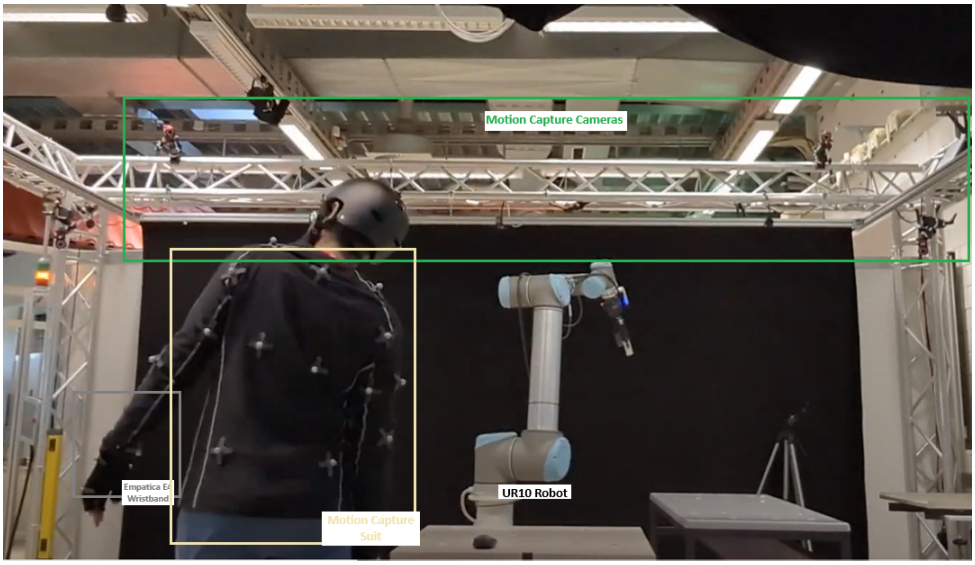


Figure 3.3: Apparatus used

For capturing the physiological signal of the participants, the Empatica E4 wristband was equipped on the participant's non-dominant hand. The participant's physiological signals such as GSR, EDA, HR and temperature, are captured by the Empatica E4 wristband, which transmits data wirelessly via Bluetooth to a Windows PC. This PC with an Intel i5-12600KF at 3.7 GHz using 32 GB RAM and a NVIDIA GeForce RTX 3060 with 12 GB VRAM. runs the E4 streaming server aswell, facilitating the real-time transfer of this data. The various motion capture cameras recording the participant's movements are synced together and are connected to the Windows PC as well running the motion capture software, Motive. The physiological data from the Empatica E4 and the motion data from the motion capture system are then streamed to a Linux PC running the Robot Operating System (ROS)1 Melodic. The motion capture data is published to */tf* topic. Whereas the Empatica E4 node is available as a ROS2 node running inside a docker container interfacing with the ROS1 using a ROS bridge. Then a data synchronization script is used to create a ROS node that subscribes to the multiple topics from various sources, synchronizes the incoming data, and publishes a compiled message to the */aggregated_data* topic. A more detailed description of the data synchronization process is provided in section 4.1. This synchronized data topic is then recorded to a rosbag. A general schematic of this is shown in Figure 6.1.

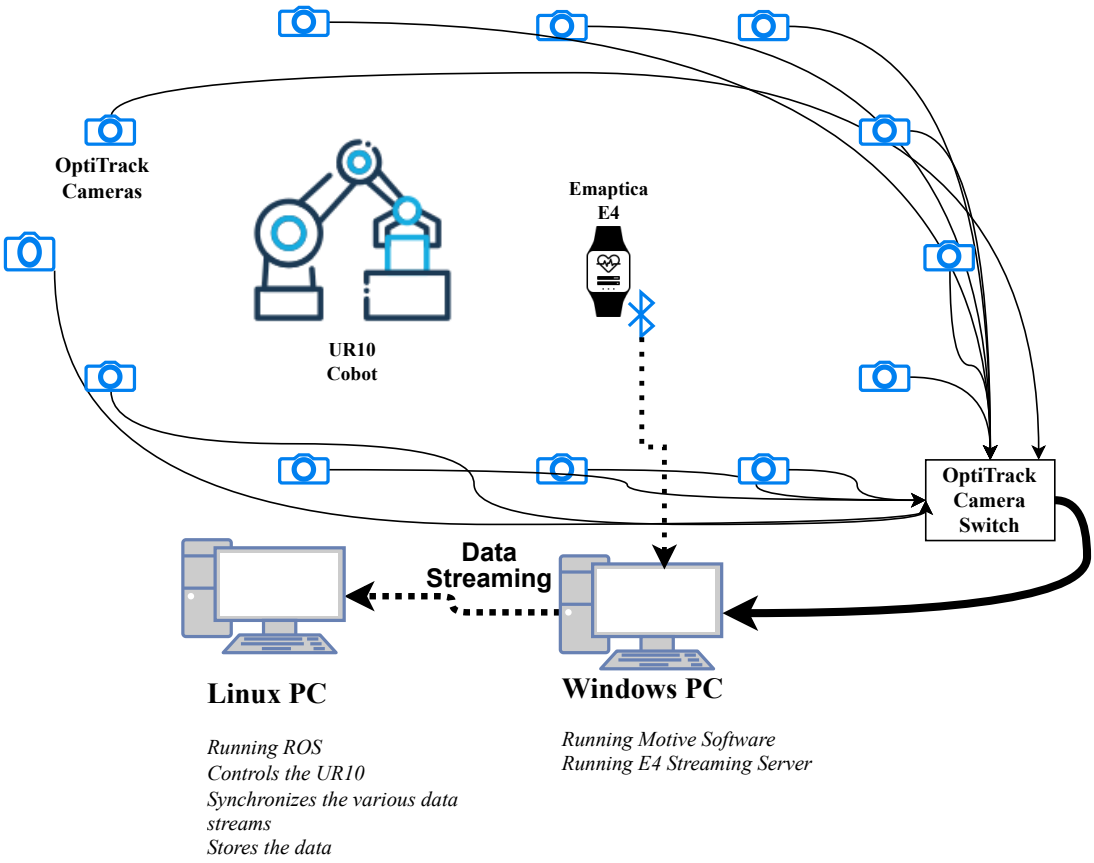


Figure 3.4: Schematic of the experiment setup

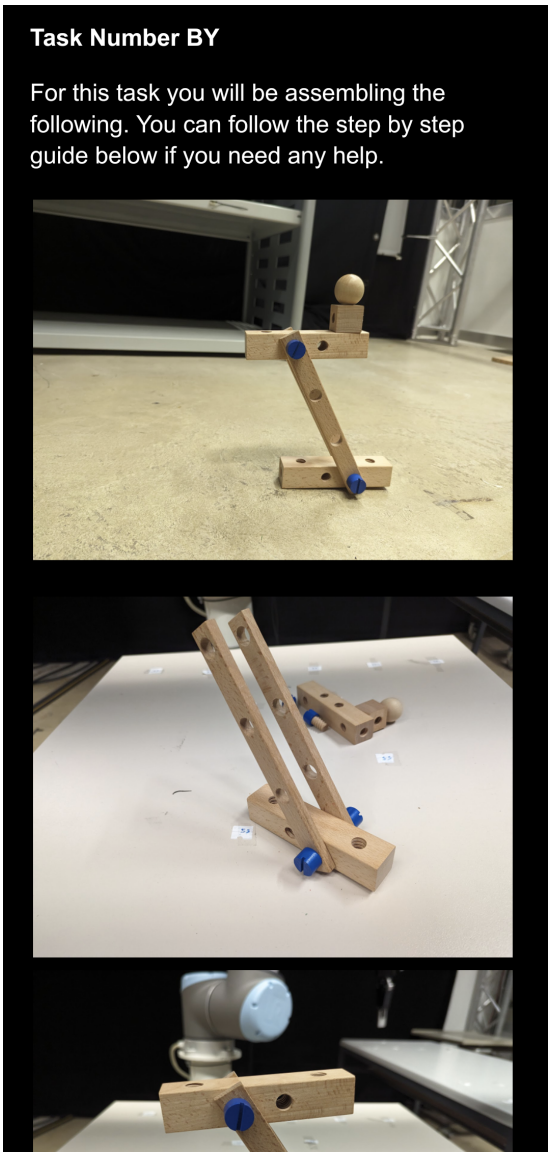
3.3 Experimental Procedure

The participants visited the experiment room in the time slot they had selected. Upon arrival, they were greeted and guided through a comprehensive orientation that explained the experimental procedures. This session included detailed descriptions of the various equipment involved, such as the Empatica E4 wristband for monitoring physiological responses and the motion capture system for observing and recording precise movement. Each participant was fitted with the motion capture suit and the Empatica E4 wristband, which was placed on their non-dominant hand, and both were carefully calibrated for accurate data collection.

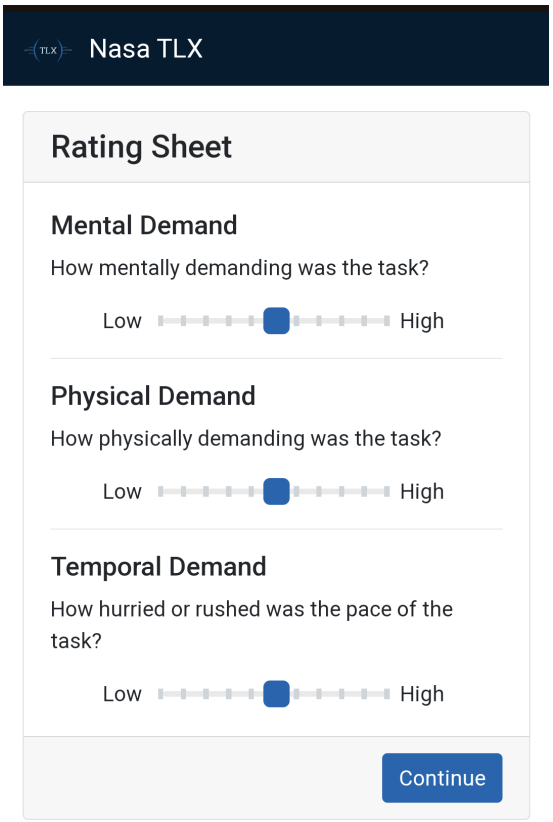
Participants then were given an initial questionnaire that included a consent form, general information, and questions about their prior experience with cobots as well as the General Attitudes Towards Robots Scale (GAToRS) questionnaire (Koverola et al. 2022).

Once the preliminary documentation was complete, we established a baseline of physiological signals for each participant, which involved recording data for 2 minutes without any interaction with the cobot. This step ensured that we had a standard reference point for each participant's physiological state prior to beginning the tasks.

The main experimental procedure involved a sequence of seven distinct tasks, with the sequence randomized for each participant to control for learning effects. Before the start of each task, a two-minute briefing was provided. This briefing not only outlined the objec-



(a) Step by Step Task instructions



(b) NASA-TLX Questionnaires

Figure 3.5: Screenshots from mobile device

tives and requirements of the task but also walked the participant through the instructions displayed on a mobile device, ensuring clarity and preparedness. Figure 3.5 illustrates the interface that participants encounter on the mobile device during the experiment. After the introduction, participants performed the task (Task i), during which both physiological and motion data were recorded. Each task lasted for about 5 minutes in average.

Upon completion of each task, participants were asked to fill out a post-task questionnaire. This included theNASA-TLX(as a webapp by Pandian and Suleri (2020)) to assess cognitive workload and the SAM to measure emotional response. These instruments were crucial for evaluating the impact of the task on the participant and infer subjective stress levels and emotional well being from each task. Whilst the participants were filling the questionnaires, the experimental setting was reset to their original position in preperation for the next task.

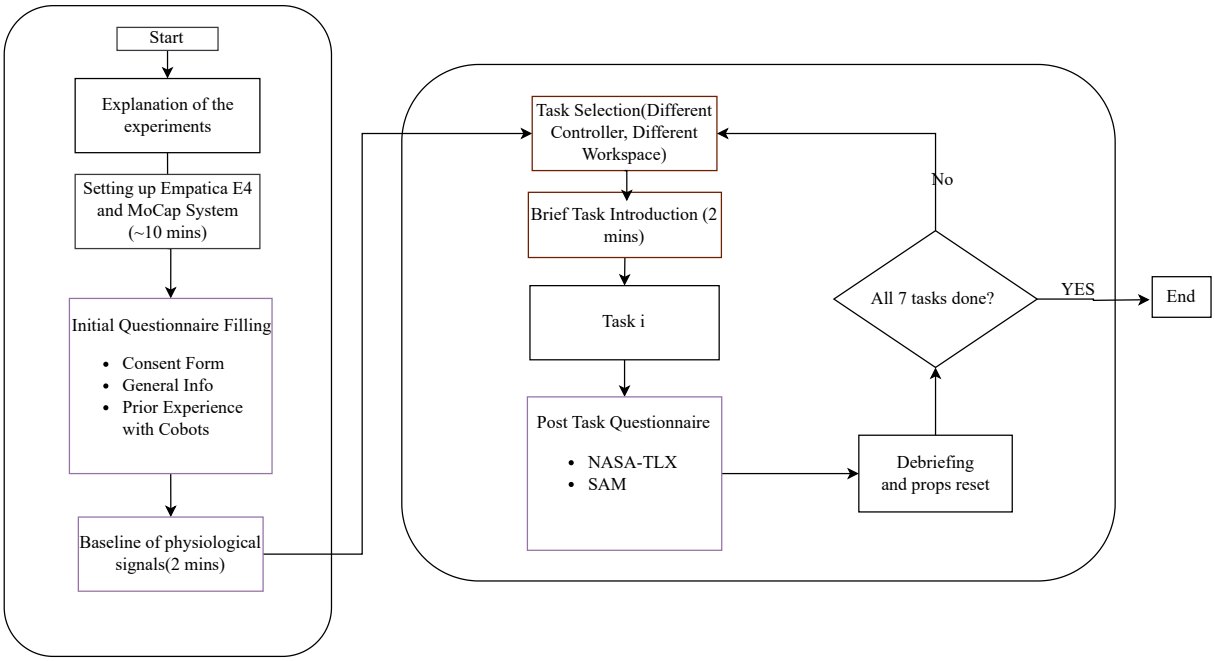


Figure 3.6: Schematic of the experiment protocol

After a participant had completed all tasks, we conducted a debriefing session. During this session, the participant could provide feedback and discuss their experiences as well explain the whole aim of the study and research. The whole session lasted for 45-60 mins on average

The structured design of this protocol ensured the collection of consistent and reliable data on human-robot interaction, with careful consideration of participant engagement and task impact.

4

Stress Detection Methodology

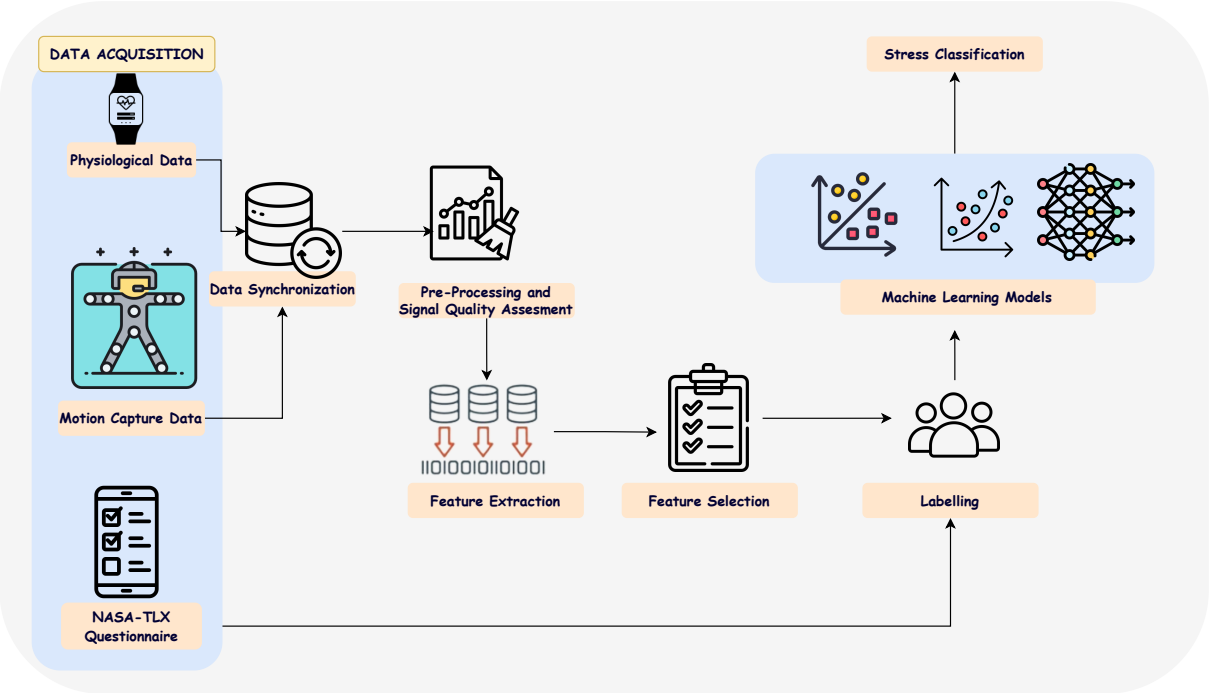


Figure 4.1: Schematic of the experiment setup

4.1 Data Synchronization‡MO

As previously eluded in section 3.2, data synchronization is crucial to our experimental framework, ensuring the consistency of data gathered from all the different sensors. In our setup, the Empatica E4 wristband collects various physiological signals, such as the Blood Volume Pulse(BVP), Electrodermal Activity (EDA), Heart Rate (HR), and Skin Temperature(ST) at different frequencies, as well as the motion capture system that tracks the participant's movements at a higher frequency. Given the varying sampling rates of these data sources, it's crucial to synchronize them to ensure they are comparable and accurate.

The Empatica E4 samples the Blood Volume Pulse (BVP) data at a rate of 64Hz. However, it transmits other metrics like EDA and temperature at a lower rate of 4Hz. On the other hand, the Motive software captures the motion capture data at a higher rate of around 120Hz. These differences in sampling rates require a careful synchronization approach.
Our synchronization strategy focuses on aligning all the different data streams to uniform and matching frequency. Considering the need for detailed data where we retain most of the information required, we standardize all data streams to a frequency of 64Hz, matching the Blood Volume Pulse (BVP) rate from the Empatica E4. This standardization process involves downsampling the motion capture data, originally recorded at a higher frequency of 120Hz. Simultaneously, for other signals like the Electrodermal Activity (EDA) and temperature, updating at a lower frequency of 4Hz, and acceleration data at 32Hz, we use a forward-filling approach, carrying their most recent values until an update occurs.
A specialized ROS node is responsible for this synchronization. By using the BVP rate as the master reference, the node ensures that both physiological and motion data are synchronized in time. This alignment is critical for an integrated and comprehensive analysis of the participant's responses, providing a dataset that accurately reflects both the physiological states and motion data.
Once synchronized, the data is then published to a ROS topic, typically */aggregated_data*. This topic carries a comprehensive stream of information that combines detailed physiological measures with motion data. This dataset forms the foundation for subsequent analysis phases in our study, such as feature extraction and stress detection.

4.2 Pre-Processing‡MO

Pre-processing is a crucial step in analysing data from physiological sensors and motion capture systems, as it prepares the raw data for subsequent processing and analysis phases.The pre-processing and subsequent feature extraction were done using the BIOBSS library in python(BIOBSS n.d.). This section outlines the key aspects of pre-processing in our study.

Data Filtering

In the pre-processing phase of our study, data filtering plays a crucial role in refining the quality of the signals gathered from the physiological sensors and motion capture systems. This step involves applying specific filtering techniques to the raw data to remove unwanted data, thereby enhancing the signal's clarity and usability for further analysis.
The primary objective of data filtering is to isolate the significant aspects of the signal while eliminating any unwanted noise or interference. Different filtering methods are employed depending on the nature of the signal and the type of noise present. For instance, we use N-th-order Butterworth filters, which effectively retain the desired frequency range while attenuating frequencies outside this range. The Butterworth filter is known for its smooth frequency response and is particularly useful in physiological signal processing, where preserving the signal's integrity is crucial.
Each signal type dictates specific filter parameters like filter order, cutoff frequencies, and filter type (lowpass, highpass, or bandpass). This careful selection ensures that the final

signal is representative of the true physiological data crucial for accurate analysis.

Data Normalization

Normalisation is a critical step in data pre-processing as well, particularly when dealing with signals of varying magnitudes or scales. Our approach involves applying a normalisation process to each input signal, standardising the data values range. This step is essential for comparing and combining data from different sensors effectively.
The method we use for normalisation is primarily the 'z-score' method. This technique transforms the data into a mean of zero and a standard deviation of one. Doing so ensures that each signal contributes equally to the analysis, irrespective of their original scale or distribution. This standardisation is crucial for machine learning models, as it enhances algorithm performance and prevents any single feature from dominating due to its scale. Normalisation also aids in mitigating the impact of outliers, as it brings all data points onto a common scale, making them more suitable for analysis.

Signal Quality Assessment

Signal quality assessment is an integral part of the preprocessing phase to ensure the reliability and accuracy of data collected from sensors, which is fundamental for accurate analysis. Various methods are employed to assess the quality of signals, each targeting specific types of anomalies or artefacts.

Detection of Clipped Segments: This method involves identifying segments in the signal that are clipped or truncated. Clipping often occurs when the signal amplitude reaches the sensor's recording capacity limits. By setting thresholds for positive and negative clipping, the method detects and marks these segments, facilitating their exclusion or correction.

Detection of Flatline Segments: This method identifies flatline segments where the signal shows minimal variation over a period. Such segments can indicate sensor displacement or malfunction. The method identifies these periods by assessing the duration of flatness and the threshold for change in signal amplitude, helping exclude non-physiological data from the analysis.
Each method plays a crucial role in verifying the integrity of the signal data. Identifying and addressing issues like clipping, flat-lining, and inconsistent patterns, signal quality assessment ensures that subsequent data analysis stages are based on accurate and reliable data.

Baseline Correction

Baseline correction forms a pivotal part of our data normalization strategy, particularly tailored for participant-specific physiological data. This approach is centered around the concept of adjusting the data relative to each participant's baseline physiological state, captured during a rest period prior to the experimental tasks. This preparatory measure establishes a reference point against which subsequent physiological responses are

compared. In the baseline correction process, we begin by computing the average values of physiological signals recorded during the baseline phase before the start of the experiment. This baseline phase is critical as it represents a period of rest where the participant's physiological state is unaffected by experimental stressors. By establishing this baseline, we are able to set a reference point that reflects the participant's normal physiological state. Subsequently, we adjust the data points collected during the active phases of the experiment relative to these baseline averages. This adjustment is a normalization process that centers the data around a personalized zero point, effectively accounting for individual physiological variations. The core advantage of baseline correction lies in its ability to mitigate the influence of inter-individual variability on the physiological measurements. Since each participant exhibits unique baseline characteristics and responds differently to stressors and other factors, the process of normalizing data against individual baselines serves as a valuable means to address this variability and achieve a more accurate and personalized assessment of stress responses.
This method ensures that the changes observed in the physiological data during the experiment are indicative of the participant's response to the experimental conditions rather than being a reflection of their baseline physiological state.

Signal Segmentation

In our research, the segmentation of physiological data into windows was a crucial part of the preprocessing. This process involved breaking down the continuous data streams into smaller, manageable sliding windows for detailed analysis. The selection of window size and step size was critical and was tailored based on the characteristics of the signal and the objectives of our analysis.
The window size was carefully chosen to capture relevant physiological and behavioral patterns within each segment, balancing the need to encapsulate meaningful data against the computational demands of processing. The step size determined the overlap between these windows, ensuring continuity and that no significant transient events were missed.
Accounting for the sampling rate of each signal was vital in customizing the segmentation process appropriately. This flexible approach was key to accommodating different types of signals, ensuring that the window size was appropriate for the length of the signal and that the segmentation parameters were compatible with each signal's nature.
Segmenting data into Windows enabled us to convert the ongoing data streams into a format suitable for comprehensive analysis. This structured approach facilitated subsequent computational processes, including feature extraction and pattern recognition, essential for robust stress detection and analysis. This method of using windows in data segmentation is fundamental in ensuring that each part of the continuous data stream is analyzed effectively, allowing for a thorough understanding of stress indicators within the dataset.

4.3 Feature Extraction ‡MO

Feature Extraction and Selection play a pivotal role in the effectiveness of machine learning models, especially in the context of human stress recognition. Feature extraction involves deriving meaningful attributes from the raw data collected. The features extracted can

vary widely, including statistical features, time-domain, frequency-domain, and linear and non-linear features.
The complexity of these features can range from basic statistical measures like mean, median, minimum, and maximum to more intricate features based on specific data modalities. Each used in stress detection may yield a unique set of features, contributing to the overall data analysis and model accuracy. The selection and application of these features are crucial, as they directly impact the classification stage, ultimately influencing the model's performance in stress recognition.
Comprehensive reviews have been conducted on this, notable Giannakakis et al. (2022), Arsalan et al. (2023). We have utilized these extensive analyses as a foundation to select and identify the appropriate features into our study.

4.3.1 BVP-Blood Volume Pressure

Photoplethysmography (PPG) sensors provide a non-invasive optical method to acquire Blood Volume Pulse (BVP) signals, detecting volumetric blood flow changes as explained 2.3.1. Among the various metrics that can be extracted from PPG signals, Heart Rate (HR) and Heart Rate Variability (HRV) are prominent. These features offer critical insights into cardiac function and stress response. Detailed discussion of HR and HRV feature extraction from PPG will follow later.
After initial pre-processing of the PPG signal, which includes filtering, baseline correction, normalization, and signal quality assessment, the pivotal step in signal analysis is peak detection. This involves accurately identifying systolic peaks in the blood volume pulse, crucial for calculating HR and HRV aswell.
The PPG waveform typically comprises two peaks: systolic and diastolic. While systolic peaks are usually prominent, diastolic peaks may not be observable in certain conditions. However, when identifiable, diastolic peaks offer additional information, contributing to a more comprehensive analysis.
To locate the diastolic peak, analysis often involves examining the first and second derivatives of the PPG signal, known as the Velocity Plethysmogram (VPG) and Acceleration Plethysmogram (APG), respectively (Suboh et al. 2022). Identifying fiducial points on VPG and APG is critical, as these points can provide insights into blood pressure estimation and other advanced cardiovascular analyses.
PPG signal analysis encompasses a variety of features across different domains:

Time Domain/Morphological Features : These features are directly extracted from the morphology (shape and structure) of the PPG waveform. These include cycle duration, peak amplitudes, and ratios of different waveform components. These features give insights into the blood volume changes with each heartbeat and can indicate changes in peripheral blood flow dynamics.

Frequency Domain Features : Analysis of the frequency components of the PPG/BVP signals reveals the rhythmic patterns linked to cardiovascular dynamics. This typically involves power spectrum analysis to identify dominant frequencies.

Statistical Features : The statistical analysis of PPG/BVP signals includes calculating mean, standard deviation, skewness, and kurtosis, offering a comprehensive statistical overview of the waveform. Advanced Feature Extraction through VPG and APG Analysis further deepens the understanding of cardiovascular dynamics. These derivatives of the PPG signal expose intricate details about blood flow, particularly regarding systolic and diastolic activities. Features derived from VPG and APG include amplitudes and durations of specific waves and ratios comparing different waveform components.

HR Features

Heart Rate (HR) is a fundamental measure in cardiovascular and stress-related studies, representing the frequency of the heartbeat.It is considered to be the most widely adopted and straightforward measure to estimate stress levels (Giannakakis et al. 2022) It is typically expressed in beats per minute (bpm). The primary method of deriving HR from PPG involves counting the number of systolic peaks within a specified time frame.
Key Features and Analysis in HR:

- **Mean and Standard Deviation of the R-R Interval:** Provide a basic understanding of heart rate variability. The mean R-R interval offers insight into average heart rate, while the standard deviation reflects the variability around this mean.
- **Root Mean Square of the Successive Differences (RMSSD):** Measures the short-term variability in R-R intervals, primarily reflecting parasympathetic nervous system activity.
- **Mean R Peak Amplitude:** The average amplitude of the R peaks in the PPG signal, indicating the strength and consistency of heartbeats.
- **Skewness and Kurtosis of R-R Intervals:** Statistical measures describing the distribution of R-R intervals. Skewness indicates asymmetry, while kurtosis indicates the 'tailedness' of the distribution.
- **Percentile of R-R Intervals:** Involves calculating specific percentiles (e.g., 50th, 95th) of the R-R interval distribution, providing additional insights into heart rate variability.

HRV Features

Heart Rate Variability (HRV) analysis, an essential aspect of cardiac function understanding, is also derivable from PPG signals. HRV refers to the variation in the time interval between heartbeats, indicated by the beat-to-beat (R-R) intervals variation. It extends beyond a mere measure of cardiac rhythm, serving as an indicator of physiological resilience and adaptability in response to stress. HRV features extracted from PPG, including R-R interval and the root mean square difference of consecutive R-R intervals, are instrumental in assessing both heart rate dynamics and autonomic nervous system regulation.
Some of the key features in HRV analysis include:

Time-Domain Features

- **SDNN (Standard Deviation of NN intervals):** Measures overall heart rate variability.
- **RMSSD (Root Mean Square of Successive Differences):** Reflects the beat-to-beat variance in heart rate and is particularly sensitive to changes in the parasympathetic nervous system.
- **NN50 and pNN50:** NN50 counts the number of pairs of successive NN intervals differing by more than 50 ms, and pNN50 is the proportion of NN50 to the total number of NN intervals.

Frequency-Domain Features

- **Low Frequency (LF):** Represents a blend of sympathetic and parasympathetic activity.
- **High Frequency (HF):** Primarily reflects parasympathetic activity.
- **LF/HF Ratio:** Used to assess the balance between sympathetic and parasympathetic nervous systems.

Non-Linear Features

- **SD1/SD2 (Poincaré Plot Analysis):** Provides a geometric representation of HRV, offering insights into the complexity of heart rate dynamics.
- **Sample Entropy:** Measures the complexity or irregularity of R-R interval time series.

4.3.2 EDA-Electrodermal Activity ‡MG

Electrodermal Activity (EDA), also known as galvanic skin response (GSR), is an indicator of emotional and physiological arousal,primarily influenced by the Sympathetic Nervous System (SNS). It primarily consists of two components: tonic (Skin Conductance Level, SCL) and phasic (Skin Conductance Response, SCR).As already explained in Section 2.3.1 the tonic component represents baseline levels of skin conductance, reflecting slow changes in arousal state. The phasic component, on the other hand, captures rapid fluctuations in response to specific stimuli or events. Dawson, Schell, and Filion (2007).
After the usual process of filtering, baseline correction and normalizing the signal as well as checking the quality of the signal we first decompose the EDA signal into its tonic and phasic components using continuous decomposition analysis. This process allows us to separately analyze the steady-state (SCL) and transient (SCR) aspects of skin conductance. The decomposition is typically carried out either using a highpass or bandpass filtering techniques or a convex optimization algorithm cvxEDA (Greco et al. 2016) , ensuring that each component accurately represents the underlying physiological processes.

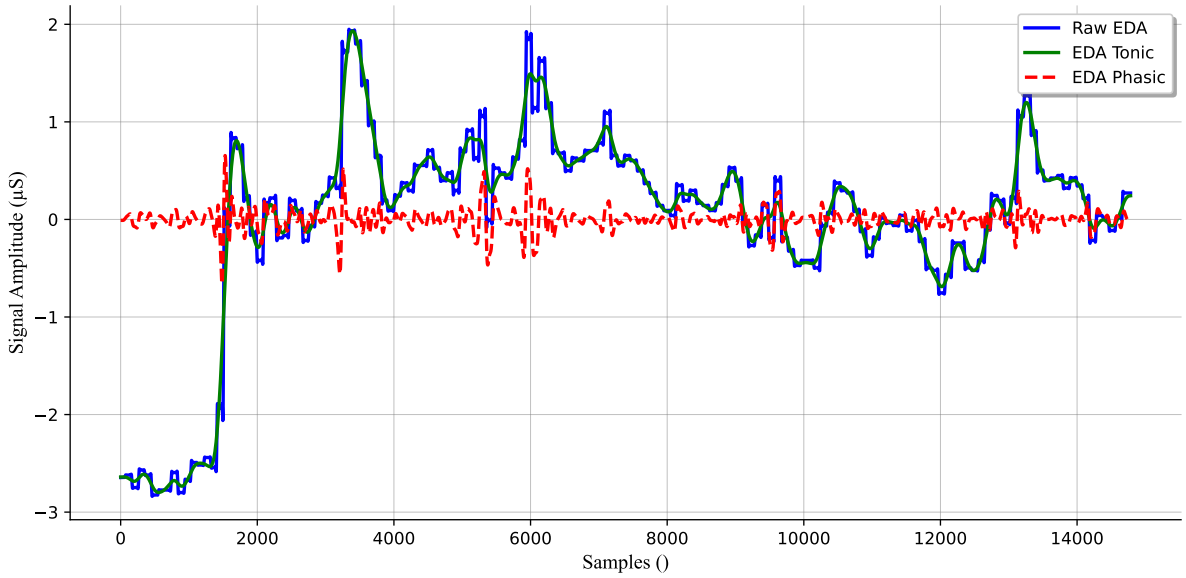


Figure 4.2: EDA

The activity of sweat glands, predominantly controlled by the SNS, leads to an increase in SCR during emotional arousal (Dawson, Schell, and Filion 2007). Notably, the non-specific SCR (NS.SCR) is intricately related to cognitive processes and psychophysiological states, acting as a direct measure of arousal (Nikula 1991). Unlike other physiological measures influenced by both sympathetic and parasympathetic nervous systems, SCR is exclusively modulated by the sympathetic nervous system, making it a reliable stress indicator (Setz et al. 2010). Under stress, both the tonic (SCL) and phasic (SCR) components escalate due to increased skin moisture (**stress**).

A thorough review and comparative analysis of various Electrodermal Activity (EDA) features have been conducted, as detailed by Shukla et al. (2021). This process involved a meticulous examination of close to 40 distinct EDA features previously identified in the literature. The outcome of this comprehensive analysis guided us in carefully selecting the features most necessary to the objectives of our research. Some of the key features from EDA include:

A detailed examination of the phasic component involves analyzing peak amplitude, frequency, and their inter-relationships in SCRs. This analysis is crucial for deciphering emotional and cognitive stress responses, as these metrics directly reflect the intensity and frequency of physiological reactions to stimuli.

From the decomposed EDA data, a variety of time domain statistical features can be extracted. These include mean (μ), standard deviation (σ), coefficient of variance (CV), variance (σ^2), and kurtosis (β) from the phasic component.

Mean (μ) provides a measure of the central tendency of the SCR amplitudes. Standard deviation (σ) and variance (σ^2) capture the variability or dispersion around the mean. The coefficient of variance (CV) offers a normalized measure of dispersion relative to the mean. Kurtosis (β) evaluates the peakedness or flatness of the distribution of SCR amplitudes.

In addition to time-domain features, we analyze the EDA signal in the frequency domain. Furthermore, frequency-domain features like spectral power in specific bands (f1sc, f2sc,

f3sc) and the overall energy and entropy of the signal gave us a spectrum-based view of the EDA responses. By analyzing these features, we could discern patterns and rhythms in the EDA that are not immediately apparent in the time-domain.
Detecting and analyzing peaks in the phasic component (SCRs) is crucial. Peak amplitude, frequency, and their inter-relationships can be strong indicators of emotional and cognitive stress responses.
By examining both tonic and phasic components, we can understand the sustained arousal level (SCL) and the specific responses to stimuli (SCR). The correlation between these components can provide valuable insights into how sustained stress levels influence responses to immediate stimuli. From the tonic component, which encapsulates the underlying level of arousal, we calculated the mean, capturing the central tendency over time, and the standard deviation, offering insights into the variability around this mean. The maximum and minimum values, along with the range, provided us with the extremes of the EDA signal, painting a picture of the breadth of responses.
From the phasic component, we focused on the Skin Conductance Responses (SCRs) to discern more rapid changes associated with specific stimuli. We extracted features like SCR amplitude, which reflects the intensity of the response, and the frequency of these SCRs, indicating how often these responses occur. The kurtosis of SCR amplitudes, a measure of the 'tailedness' of the distribution, gave us an understanding of how peaked or flat the distribution of responses was, while the skewness indicated any asymmetry, offering clues about the predominant direction of the response distribution.
We also looked at the root mean square (SCR RMS), which is a measure of the signal's magnitude, providing a summative measure of the signal's complexity over a given period. The integral of the SCR signal (SCR Integral) was calculated to understand the total magnitude of these phasic responses over time. These features, alongside others like SCR momentum, which is akin to the second moment of the distribution, provided a comprehensive statistical breakdown of the phasic EDA signals.

4.3.3 Body Features ‡MO

Since we captured human motion using the motion capture system, we selected 13 key points from the 25 marker points used by the system, focusing on the upper body. The chosen points were Hip, Ab, Chest, Neck, Head, Left Shoulder (LShoulder), Left Upper Arm (LUArm), Left Forearm (LFArm), Left Hand (LHand), Right Shoulder (RShoulder), Right Upper Arm (RUArm), Right Forearm (RFArm), and Right Hand (RHand). These points were strategically selected to comprehensively capture the whole upper body movements. The arrangement of these points is depicted in Fig 4.3

Self Touching

Giakoumis et al. (2012) has shown that body posture and body language can be valuable indicators of stress. In line with this, we also explored body language cues such as self-touching, which Harrigan (1985)suggests can be indicative of negative affect, such as anxiety or discomfort. Specifically, we focused on face and head touching as potential stress indicators.

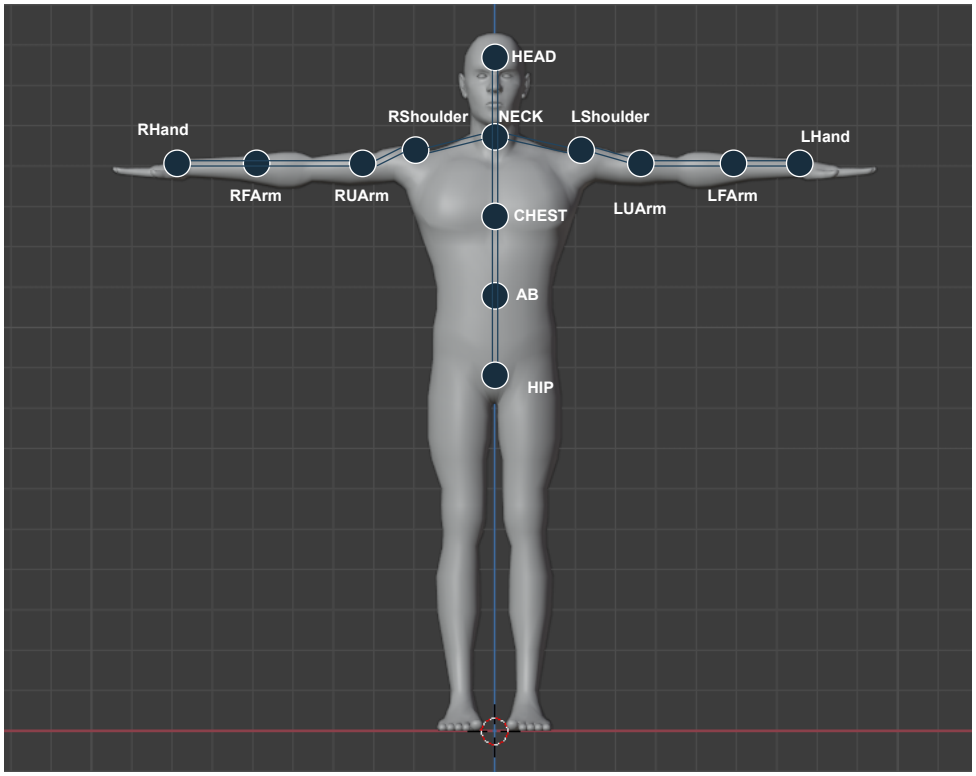


Figure 4.3: 13 points from the motion capture

To determine whether an individual is touching their face, the distances between their hand and head, and between their hand and neck, are measured. When either of these distances falls below a specified threshold, it's interpreted as a face-touching event. We utilize the frequency of these face-touching counts (FTC) and the mean duration of each occurrence (FTMD) as features.(Aigrain 2016)
To determine gestures such as face touching, */tf* data, which typically includes the position and orientation of each joint in space, can be utilized to calculate the distance between any two points. For example, if you have the coordinates of the right hand (`RHand`) and the head (`HEAD`), you can compute the Euclidean distance between these two points at each time frame to detect when the hand is close enough to the head to indicate potential face touching.
Let's define the 3D coordinates for the right hand and head at time t as:

- $P_{\text{RHand}}(t)$ for the position of the right hand at time t, with coordinates $(x_{\text{RHand}}(t), y_{\text{RHand}}(t), z_{\text{RHand}}$
- $P_{\text{HEAD}}(t)$ for the position of the head at time t, with coordinates $(x_{\text{HEAD}}(t), y_{\text{HEAD}}(t), z_{\text{HEAD}}(t))$.

The distance between the right hand and the head at time t is then calculated with the formula:

$$D_{\text{RHand-HEAD}}(t) = \sqrt{(x_{\text{RHand}}(t) - x_{\text{HEAD}}(t))^2 + (y_{\text{RHand}}(t) - y_{\text{HEAD}}(t))^2 + (z_{\text{RHand}}(t) - z_{\text{HEAD}}(t))^2} \tag{4.3.1}$$

If this distance, $D_{\text{RHand-HEAD}}(t)$, is less than a certain threshold, denoted as θ, it suggests that the right hand is in proximity to the head, indicating potential face touching.

To determine the occurrence of face touching, you would track when this distance becomes less than the threshold θ and also ensure that the hand remains within this threshold for a certain duration to count as an occurrence. The distance for the left hand can be calculated in a similar manner.
To compute the number of occurrences (*FTC-Face Touching Count*) and the average duration (*FTMD-Face Touching Mean Duration*) of face touching is shown in Algorithm 4.3.1.

Algorithm 4.3.1.: Face Touching Detection

Require: Set of time-stamped positions from the `/tf` topic, Threshold distance θ

function DetectFaceTouching
  Initialize $FTC \leftarrow 0$ ▷ Occurrences of face touching
  Initialize $TotalDuration \leftarrow 0$ ▷ Total duration of face touching events
  Initialize $FTMD \leftarrow 0$ ▷ Mean duration of face touching events
  for each time stamp t_i in `/tf` topic **do**
    Calculate $D_{RHand-Head}(t_i)$ and $D_{LHand-Head}(t_i)$
    if $D_{RHand-Head}(t_i) < \theta$ **or** $D_{LHand-Head}(t_i) < \theta$ **then**
      Start duration counter
      $FTC \leftarrow FTC + 1$
    if either distance exceeds θ **then**
      Stop duration counter
      Add duration to $TotalDuration$
  if $FTC > 0$ **then**
    $FTMD \leftarrow \frac{TotalDuration}{FTC}$ ▷ Calculate mean duration
  return FTC, $FTMD$

Sudden Movement

Lagomarsino et al. (2022) suggests another way in which motion data can be used to assess stress that is by identifing periods of high activity/sudden activity called hyperactivity. Sudden movements, or abrupt changes in body motion, can be indicative of stress responses. These movements are characterized by significant deviations from a person's regular movement patterns and can be quantitatively assessed using motion capture data. The following equations describe the computational process used to analyze sudden movements and infer stress.

$$m_j^k = \sum_{i=0}^{\tau-1} d_j^{k-i,k-i-1} \tag{4.3.2}$$

Equation 4.3.2 defines the movement of the j^{th} joint within a time window τ as the sum of the displacements between consecutive frames.

$$\Delta_j^k = m_j^k - \mu_j \tag{4.3.3}$$

In Equation 4.3.3, Δ_j^k represents the deviation of the j^{th} joint's movement from its baseline mean motion μ_j, calculated during a calibration phase.

$$a_j^k = \begin{cases} \frac{\Delta_j^k}{\sigma_j} - 1 & \text{if } \Delta_j^k > \sigma_j \\ 0 & \text{otherwise} \end{cases} \tag{4.3.4}$$

Equation 4.3.4 assesses the activity level a_j^k for the j^{th} joint, taking into account the standard deviation σ_j as a threshold for sudden movement.

$$a_k = \min\left(1, \frac{1}{N}\sum_{j=1}^{N} a_j^k\right) \tag{4.3.5}$$

Finally, Equation 4.3.5 calculates the overall level of sudden movement at time instance k by averaging the activities across all joints, thus providing a descriptor of hyperactivity or sudden movement.
This method allows for a comprehensive analysis of the motion data to identify periods of high activity that may correlate with stress responses.
While our study focuses on the analysis of upper body movements for stress detection, it is important to note that other bodily cues can also be significant indicators of stress or anxiety. One such example is the rapid tapping or bouncing of one's feet, which is often a subconscious response to nervous energy or unease. Such movements are typically a form of self-soothing behavior that occurs when an individual is experiencing discomfort or stress.
Unfortunately, due to the scope of our study setup, we restricted our tracking to the upper body and therefore could not capture lower body movements, such as foot tapping or leg bouncing. These actions could potentially provide additional insights into a participant's stress levels and offer a more comprehensive understanding of physical stress responses.
Including lower body data in future studies could enhance the detection and analysis of stress indicators, allowing for a fuller picture of the physiological and behavioral state of an individual under stress. This would enable us to capture a wider range of stress-related behaviors and potentially increase the accuracy and reliability of stress detection in real-time scenarios.

4.4 Feature Selection

From the various number of features that are extracted from the signals in the previous sections, it is important to select only a few features for several reasons. The primary reason for this selective approach is to enhance the model's generalizability. Models that are trained using an excessive number of features, especially those that are less significant or redundant, have a tendency to overfit the training data. This means that they capture random variations instead of the actual underlying patterns, resulting in poor performance when applied to new, unseen data. Moreover, a smaller number of features mitigates the risk of collinearity, where interdependent variables can distort the model's predictive power. It also reduces computational load, leading to more efficient models.

Initially, we utilize a correlation matrix for analyzing our various extracted features. This matrix is important in revealing the relationships and interdependencies among different features. Mainly, we focus on identifying features highly correlated with the target variable, stress levels, while exhibiting low inter-correlation with each other. Such features are likely to carry unique information beneficial for our model. On the other hand, features that show a strong correlation with each other, meaning they share similar information, could provide redundant data. In such cases, it might be beneficial to keep only one feature from each correlated pair to minimize redundancy and avoid overfitting of the model.
In addition to using the correlation matrix for feature selection, we thoroughly examine the current literature on stress detection and physiological research. Examining previous literature helps to confirm the results and ensures that our selection of features does not exclude important elements that have been shown to be successful in earlier studies. The literature review plays a crucial role as it connects our real-world observations with existing research, adding another level of confirmation to our process of selecting features. It helps to validate our choices by aligning them with existing knowledge and understanding in the field. As a first evaluation only HRV, HR and EDA features are used and Table 4.1 shows the list of features that were selected for the model. Appendix shows the whole list of possible features that can be extracted from the data.

4.5 Ground Truth Labeling[‡MO]

In supervised machine learning, having a reliable dataset with accurately labeled data is crucial for developing effective models. Labelling involves assigning meaningful labels to data instances. These labels serve as the ground truth against which model predictions are evaluated. Ground truth labeling, especially in the context of stress detection, is a challenging task especially due to the multifaceted nature of the data.
In existing stress measurement literature, several methods have been employed to label stress levels. One common approach involves designing experiments that deliberately induce stress through specific tasks or conditions, such as the Stroop test(Stroop 1935) or the Trier Social Stress Test (TSST) (Alshamrani 2021; Kraaij, Koldijk, and Sappelli 2014; Schmidt et al. 2018; Smets et al. 2018). These controlled scenarios create environments where stress levels can be objectively measured and labeled. Another prevalent method is the use of third-party observation, where an external observer assesses and numerically scores the subjects responses to certain situations, thus determining the level of stress experienced (Aigrain 2016; Jin, Osotsi, and Oravecz 2020; Siirtola and Röning 2020).Another method is where people use biosignals which are assumed to directly quantify stress, and deviation from the baseline values of the biosignals are taken as moments of stress. Kyriakou et al. (2019) developed a rule based algorithm with defined threshold using variations to the galvanic skin response and skin temperature as direct indicator of moments of stress.
We utilized another common technique using self-reporting, where participants directly assess their own stress levels and report them. For this purpose, we employed the NASA Task Load Index NASA-TLX method, a widely recognized tool for gauging subjective workload and stress as explained in section 2.2. We employed the NASA-TLX to assess

Feature	Description	Signal Type	Domain	Expected Behavior
eda_std	Standard deviation of Electrodermal Activity	EDA	Time Domain	Increase
eda_range	Range (difference between maximum and minimum) of EDA	EDA	Time Domain	Increase
scr_std	Standard deviation of Skin Conductance Response	EDA	Time Domain	Increase
scr_range	Range (difference between maximum and minimum) of SCR	EDA	Time Domain	Increase
eda_kurtosis	Kurtosis of Electrodermal Activity	EDA	Time Domain	Varies
scr_rms	Root mean square of Skin Conductance Response	EDA	Time Domain	Increase
scr_integral	Integral (area under the curve) of Skin Conductance Response	EDA	Time Domain	Increase
scl_std	Standard deviation of Skin Conductance Level	EDA	Time Domain	Increase
hrv_mean_nni	Mean of NN intervals in Heart Rate Variability	HRV	Time Domain	Decrease
hrv_median_nni	Median of NN intervals in Heart Rate Variability	HRV	Time Domain	Decrease
hrv_sdnn	Standard deviation of NN intervals	HRV	Time Domain	Decrease
hrv_rmssd	Root mean square of successive NN interval differences	HRV	Time Domain	Decrease
hrv_mean_hr	Mean Heart Rate in Heart Rate Variability	HRV	Time Domain	Decrease
hrv_mad_nni	Mean absolute deviation of NN intervals	HRV	Time Domain	Varies
hrv_SampEn	Sample Entropy in Heart Rate Variability	HRV	Non-Linear Domain	Varies

Table 4.1: Features Relevant to Stress Detection

participants' subjective experiences of stress and to establish a ground truth for their stress levels. The increase in NASA-TLX has been widely shown to demonstrate a positive correlation with mental fatigue and stress as shown by (Nguyen and Zeng 2017; Kaduk, A. P. J. Roberts, and Stanton 2021). Bakhsh et al. (2019) documented the relationship between task load and stress among surgery residents, highlighting the use of the NASA-Task Load Index. They found a positive correlation between the NASA-TLX scores and objective stress levels, which were measured by indicators of sympathetic activity such as heart rate and blood pressure. This correlation suggests that higher task loads, as perceived by the residents, were associated with increased physiological markers of stress. Supporting these observations, Favre-Félix et al. (2022) confirmed these results using a similar method by taking acceptable thresholds and percentile ranks for the NASA-TLX scale as suggested by Grier (2015). In line with these findings, we considered an increase in NASA-TLX ratings to be an indicator of increased stress levels.
To categorize stress into three distinct levels: 'Not Stressed (0)', 'Slightly Stressed (1)', and 'Stressed (2)', we opted for the utilization of z-scores derived from the NASA-TLX weighted ratings. The choice of z-scores was driven by our aim to standardize responses across a diverse participant pool, thereby facilitating uniform comparisons and addressing variations in individual interpretations of the NASA-TLX scale.
Using z-scores offers several advantages in our context. It provides standardization, which is beneficial when dealing with varied interpretations of the NASA-TLX scale by different participants. This approach normalizes the data, simplifying the categorization into distinct stress levels based on statistical criteria. In implementing this approach, we have set specific thresholds within the z-score distribution to categorize stress levels. These thresholds were carefully chosen based on a combination of empirical evidence and insights drawn from the literature. A z-score below 0 indicates a 'Not Stressed' state, scores between 0 and 1 correspond to a 'Slightly Stressed' condition, and scores above 1 are categorized as 'Stressed'. Therefore, the labels, which constitute a significant element of our model, were assigned by solely relying on subjective self-reports to categorize stress levels. The cross-validation of this methodology, which depends entirely on self-reports for stress labeling, with physiological signals and the broader implications and the applicability of this approach is detailed in later in sec..... For the present, this self-report-based system is the cornerstone of our labeling strategy.

4.6 Classification -Stress Detection[‡MG]

In our stress detection model, we employed five machine learning algorithms with key settings for each:

- **SVM (Support Vector Machine)**: Utilized via `sklearn.svm.SVC`, exploring various kernels such as linear, RBF, polynomial, and sigmoid. Key parameters included `C=1.0` and `gamma=0.1`.

- **Random Forest**: Implemented using `sklearn.ensemble.RandomForestClassifier` with a focus on tuning parameters like the number of trees (n_estimators=100). Features were scaled using `StandardScaler`.

- **Naive Bayes**: Employed using `GaussianNB` from `sklearn.naive_bayes`, which is efficient for high-dimensional data.

- **K-Nearest Neighbors (KNN)**: Configured using `sklearn.neighbors.KNeighborsCla` with the number of neighbors set to 10 and employing a suitable distance metric.

- **AdaBoost**: Applied through `sklearn.ensemble.AdaBoostClassifier`, optimizing the number of estimators and learning rate for performance enhancement.

- **Neural Network (MLP)**: Developed with `MLPClassifier` from `sklearn.neural_network`. The model had a single hidden layer with 100 neurons, used the ReLU activation function, and employed the Adam solver with `alpha=0.0008` and `max_iter=200`.

These algorithms were chosen for their effectiveness and suitability in handling the complexities involved in stress classification. Each model was rigorously trained and evaluated using a 10-fold cross-validation approach to ensure robustness and reliability of the results.
Cross Validation techniques

Performance Assessment Metrics:

To evaluate the performance of classification models, particularly in multi-class scenarios, we use a confusion matrix and several key metrics. A confusion matrix provides a detailed breakdown of a model's predictions against the actual values, offering insights into its performance across different classes.

Confusion Matrix for a 3-Class Classification:

A confusion matrix for a 3-class classification can be represented as follows:

Actual \Predicted	**Class 1**	**Class 2**	**Class 3**
Class 1	TP_1	FP_{12}	FP_{13}
Class 2	FP_{21}	TP_2	FP_{23}
Class 3	FP_{31}	FP_{32}	TP_3

In this matrix:

- **TP (True Positive)**: Correct predictions of a specific class.

- **FP (False Positive)**: Incorrect predictions as a specific class, which actually belong to another class.

- **TN (True Negative)**: Correct predictions of instances not belonging to a specific class.

- **FN (False Negative)**: Incorrect predictions of instances as not belonging to a specific class, which actually do.

In this matrix, the diagonal elements (TP_1, TP_2, TP_3) represent the number of correct predictions for each class, while the off-diagonal elements indicate the misclassifications. Using this matrix, the following metrics are calculated:

- **Accuracy**: The proportion of correct predictions among the total number of cases examined.

$$\text{Accuracy} = \frac{\text{Sum of Diagonal Elements (TP and TN)}}{\text{Total Number of Predictions}}$$

- **Precision** : The ratio of correct positive predictions to the total predicted as that class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall (Sensitivity)** : The proportion of actual positives of a class correctly identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

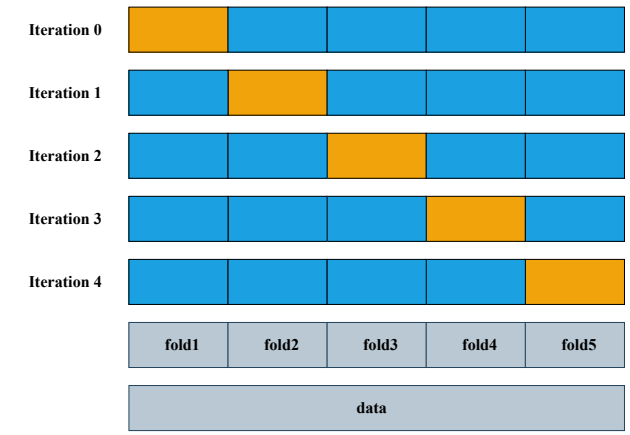
- **F1 Score** : The harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

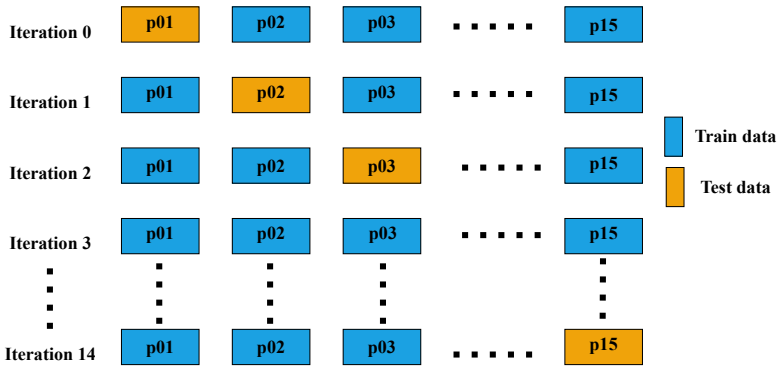
Another metric that is considered is the **Area Under the Curve (AUC)** metric of the ROC (Receiver Operating Characteristic) curve. The ROC curve plots the true positive rate against the false positive rate at various threshold settings. AUC measures the entire area underneath the ROC curve and provides an aggregate measure of the model's performance across all possible classification thresholds.
Accuracy provides an overall effectiveness of the model, while precision, recall, and F1 score offer class-specific performance insights, crucial in multi-class classification tasks.
Cross-Validation Techniques
We utilized two key cross-validation techniques to evaluate the performance and robustness of our machine learning models in stress detection: K-Fold Cross-Validation and Leave-One-Participant-Out (LOPO) Cross-Validation.

- **K-Fold Cross-Validation:**
  K-Fold Cross-Validation is a standard method for assessing the efficacy of machine learning models. It involves dividing the entire dataset into 'K' equal-sized parts, or folds. The model is trained on 'K-1' of these folds, while the remaining fold is used for testing. This process is repeated 'K' times, each time with a different fold used as the test set, ensuring that each data point is used for both training and testing exactly once. The average performance across all 'K' iterations is used to estimate the model's effectiveness.

- **Leave-One-Participant-Out (LOPO):**
  Leave-One-Participant-Out (LOPO) Cross Validation is particularly relevant for our study, which involves data collected from multiple participants. In this approach, the model is trained on data from all participants except one and then tested on the data from the left-out participant. This procedure is repeated for each participant, so in our case, with 15 participants, the process iterates 15 times. LOPO is especially beneficial in situations where participants undertake experiments in different orders.



(a) KFold



(b) LOPO

Figure 4.4: Cross Validation Techniques

By training and testing the model on data from different participants, LOPO helps minimize potential biases that can arise due to variations in how participants respond to the experiment. This method allows for an unbiased assessment of the model's performance, as each subset of data (corresponding to each participant) is tested separately, and the results are aggregated to provide a comprehensive evaluation of the model's effectiveness across the entire cohort.

Both these cross-validation methods are integral to our analysis. They provide a thorough assessment of how well our model can generalize to new, unseen data, which is crucial for developing a reliable tool for stress detection.

5 Result

5.1 Assessment of Human Stress Levels

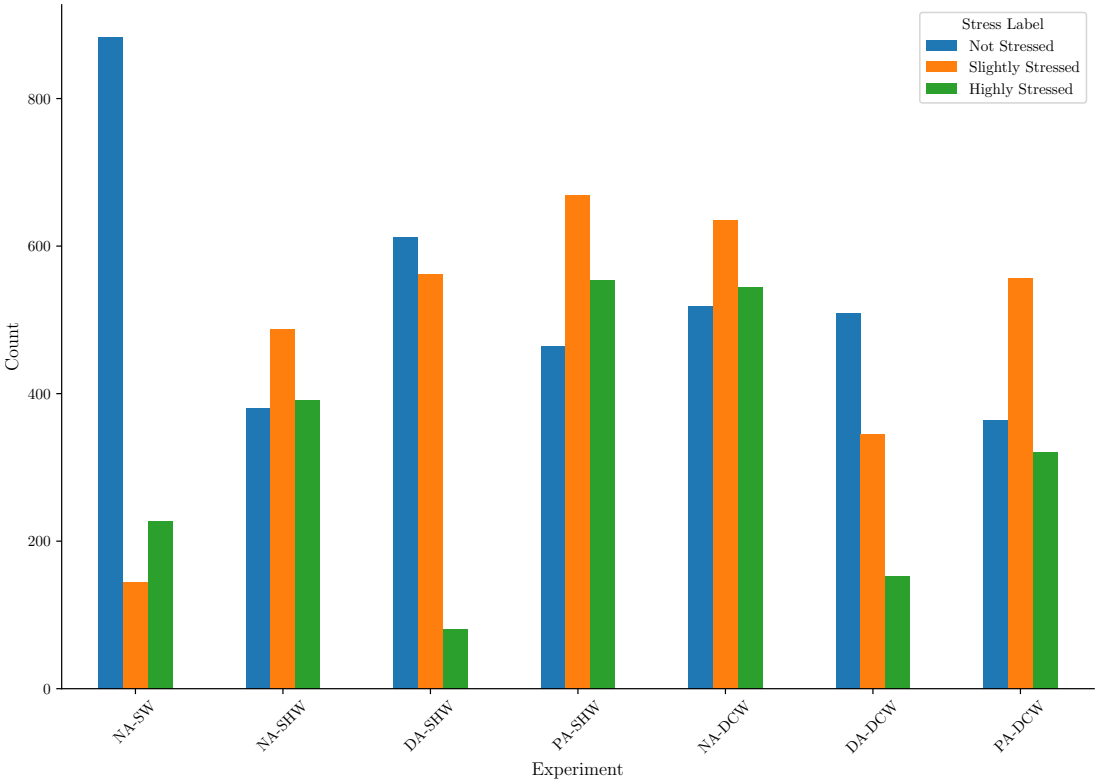


Figure 5.1: Stress levels by experiment

In this section, we present a comprehensive analysis of the perceived stress levels experienced by human participants in various our human-robot interaction scenarios. Figure 5.1 shows the 3 different stress levels experienced by participants in each of the 7 different experimental setups. Although in previous sections of our study we have used abbreviations such as AX, BX, and CX to refer to different experiment setups to match the protocol used for the data collection, for ease of understanding while presenting the results in this section we will adopt a more descriptive notation:

- **No Collision Avoidance in a Separated Workspace (NA-SW)**

- **No Collision Avoidance in a Shared Workspace (NA-SHW)**
- **Dynamic Collision Avoidance in a Shared Workspace (DA-SHW)**
- **Predictive Collision Avoidance in a Shared Workspace (PA-SHW)**
- **No Collision Avoidance with Direct Collaboration in a Shared Workspace (NA-DCW)**
- **Dynamic Collision Avoidance with Direct Collaboration in a Shared Workspace (DA-DCW)**
- **Predictive Collision Avoidance with Direct Collaboration in a Shared Workspace (PA-DCW)**

In the No Collision Avoidance in a Separated Workspace (NA-SW) scenario, the high prevalence of 'Not Stressed' responses is consistent with expectations that humans are most comfortable when working in an area separate from robots. Yet, the observation that 'High Stress' responses exceed those in scenarios with Dynamic Collision Avoidance in a Shared Workspace (DA-SHW) and Dynamic Collision Avoidance with Direct Collaboration in a Shared Workspace (DA-DCW) may initially appear contradictory. A potential rationale for this will be discussed subsequently.

For the No Collision Avoidance in a Shared Workspace (NA-SHW), there's a marked increase in stress levels, both 'Slightly Stressed' and 'Highly Stressed,' compared to the separated workspace (NA-SW). This is anticipated as sharing a workspace with robots, in the absence of any collision avoidance mechanisms, could lead to increased safety concerns and unpredictability, thereby elevating stress levels.

Another common pattern that can be seen is that introduction of Dynamic (DA) and Predictive (PA) Collision Avoidance strategies in both Shared Workspace (SHW) and Direct Collaboration (DCW) scenarios has generally led to a reduction in stress levels, with a particularly noticeable decrease in the DCW scenarios. However, this trend is not as apparent at first in the Predictive Collision Avoidance in a Shared Workspace (PA-SHW) scenario, where the number of 'Highly Stressed' instances is unexpectedly higher than in the scenario without collision avoidance (NA). It is important to consider that the overall count in the PA-SHW and No Collision Avoidance with Direct Collaboration in a Shared Workspace (NA-DCW) scenarios is higher than in other contexts, which may have influenced the stress results. The data shows the following count of instances for each task NA-SW: 1254, NA-SHW: 1258, DA-SHW: 1255, PA-SHW: 1687, NA-DCW: 1698, DA-DCW: 1007, PA-DCW: 1242. This discrepancy suggests that tasks PA-SHW and NA-DCW required more time for completion, potentially skewing the perceived stress levels.

Efforts were made to standardize each task's difficulty and completion time, but the tasks involving Predictive Collision Avoidance in a Shared Workspace (PA-SHW) and No Collision Avoidance with Direct Collaboration in a Shared Workspace (NA-DCW) inherently took longer, likely due to the task design that included attaching wheels to the base items at four separate locations.(Refer Figure 3.1).

Interestingly, another experiment that involved attaching wheels to base items was NA-SW. However, because there was no delay for the human to wait for the robot to deliver parts, this may have counteracted the longer task time. Nevertheless, the requirement of attaching wheels might have contributed to increased frustration levels, potentially explaining the slightly higher count of 'Highly Stressed' in the NA-SW scenario.

Interestingly, the tasks with the fewest attachment points—Dynamic Collision Avoidance in a Shared Workspace (DA-SHW) with five, and Dynamic Collision Avoidance with Direct Collaboration in a Shared Workspace (DA-DCW) with four—recorded the lowest counts of 'Highly Stressed' responses.

The diversity in task design was intentional to minimize the learning effects that could arise from performing similar tasks repeatedly. The goal was to create tasks that were distinct yet comparable in difficulty. The maximum number of attachment points was seven, with the minimum being four. However, it appears that this approach may have inadvertently introduced a variable of task complexity that was not adequately accounted for in the study's design.

Some other clear patterns shown are the increase of stress levels from the Dynamic collision avoidance (DA) to the Predictive collision avoidance (PA) in both the Shared workspace (SHW) and the Direct Collaboration (DCW) scenarios.

A potential explanation for this pattern is that the Predictive Collision Avoidance system may not be finely calibrated for the range of human actions during the item handover task. This lack of precise paramter tuning leads to unpredictability in the robot's actions, which are based on future projections. The added complexity of the robot's behavior can contribute to higher stress levels for the human workers.

During the experiments, variations in human behavior were observed. Some participants simply extended their hands to the robot for the handover, while others leaned in with their entire body, prompting the robot to misinterpret the action as a potential collision and retreat from the handover point for safety reasons. This forces the human to revert to their original position and attempt the handover again, leading to potential frustration and increased stress.

It's important to note that the optimization of the Predictive Collision Avoidance system's parameters was conducted prior to data collection, considering only a single standard behavior rather than the diverse behaviors of different participants. Future work could focus on refining the system to accommodate a wider range of human behaviors during handover tasks. Additionally, a hybrid approach that utilizes Predictive Collision Avoidance for general tasks and switches to Dynamic Collision Avoidance for handovers is proposed as a potential solution to reduce stress. This approach is further discussed in the **??**.

5.2 Machine Learning Classification Models

In the evaluation of stress detection models, several machine learning algorithms were compared.

Support Vector Machine

Specifically, Support Vector Machine (SVM) models with various kernel functions were employed. For the SVM with Linear Kernel, a linear kernel was utilized with a regularization parameter (C) set to 1.0, enabling probability estimation. The SVM with Polynomial Kernel employed a polynomial kernel of degree 3, also with a regularization parameter of 1.0 and probability estimation enabled. Additionally, the SVM with Radial Basis Function (RBF) Kernel utilized an RBF kernel with a regularization parameter of 1.0 and a gamma value of 0.1 for the kernel coefficient, allowing for probability estimation. Lastly, the SVM with Sigmoid Kernel employed a sigmoid kernel with a regularization parameter of 1.0 and a coefficient of 0.0. All models underwent evaluation using a 10-fold cross-validation technique. Detailed results, including mean scores and confusion matrices, can be found in Table 5.1 and Figures 1-4, respectively.

Model	Accuracy	F1 Score	Precision	Recall	AUC
SVM with Linear Kernel	0.628	0.622	0.630	0.628	0.805
SVM with Polynomial Kernel	0.750	0.745	0.803	0.750	0.946
SVM with RBF Kernel	**0.916**	**0.915**	**0.926**	**0.916**	**0.994**
SVM with Sigmoid Kernel	0.445	0.440	0.449	0.445	0.599

Table 5.1: Mean Scores for SVM Models with K-Fold Cross-Validation ($k = 10$)

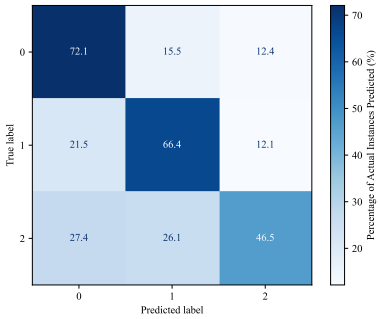


Figure 5.2: Confusion Matrix for Linear SVM

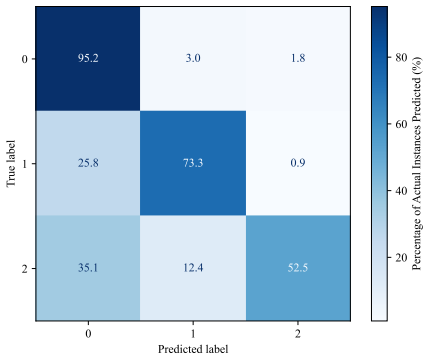


Figure 5.3: Confusion Matrix for Polynomial SVM

Among the models, the SVM with RBF Kernel achieved the highest mean scores across all metrics, including AUC, accuracy, F1 score, precision, and recall. This performance may be attributed to the RBF kernel's ability to capture non-linear relationships in the data effectively.

Random Forest

The results obtained from Random Forest models with different numbers of trees (n_-estimators) are shown in Table 5.2
This table presen

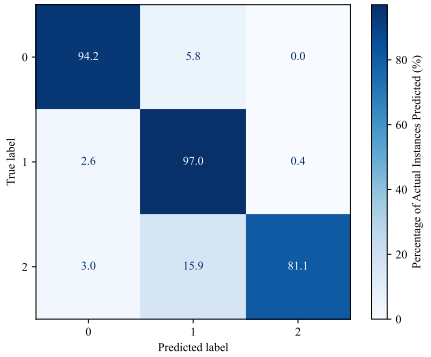


Figure 5.4: Confusion Matrix for RBF SVM

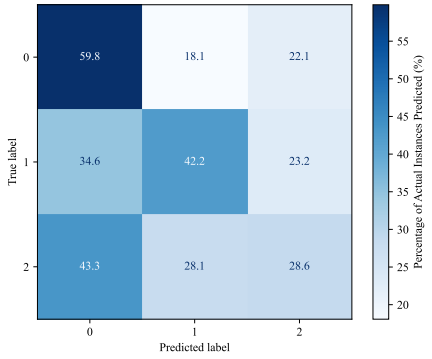


Figure 5.5: Confusion Matrix for Sigmoid SVM

Model	Accuracy	F1 Score	Precision	Recall	AUC
Random Forest (n_estimators=100)	0.949	0.949	0.950	0.949	0.992
Random Forest (n_estimators=200)	0.949	0.949	0.950	0.949	0.993

Table 5.2: Mean Scores for Random Forest Models with Different Numbers of Estimators

Naive Bayes

The Gaussian Naive Bayes classifier, with its assumption of feature independence, yielded moderate performance across various metrics, as shown in Table 5.3.

Model	Accuracy	F1 Score	Precision	Recall	AUC
Gaussian Naive Bayes	0.532	0.507	0.550	0.532	0.702

Table 5.3: Mean Scores for Gaussian Naive Bayes with K-Fold Cross-Validation ($k = 10$)

k-Nearest Neighbors (k-NN)

In this study, the k-Nearest Neighbors (k-NN) algorithm was utilized with varying values of the parameter k. The default distance metric used by k-NN, Euclidean distance, was employed for calculating the proximity between data points. Four different configurations of the k-NN model were examined, with k values set to 5, 10, 15, and 20, respectively. Each k-NN model underwent a 10-fold cross-validation procedure for evaluation. The results, including mean scores and confusion matrices, are presented in Table 5.4 and Figures 5-8, respectively.

AdaBoost

The AdaBoost models were trained using the AdaBoost classifier with varying numbers of estimators. Specifically, the models were configured with 50, 100, 150, and 200 estimators.

Model	Accuracy	F1 Score	Precision	Recall	AUC
k-NN (k=5)	**0.909**	**0.909**	**0.913**	**0.909**	**0.982**
k-NN (k=10)	0.833	0.833	0.844	0.833	0.954
k-NN (k=15)	0.780	0.779	0.795	0.780	0.924
k-NN (k=20)	0.780	0.779	0.795	0.780	0.924

Table 5.4: Mean Scores for k-NN Models with K-Fold Cross-Validation ($k = 10$)

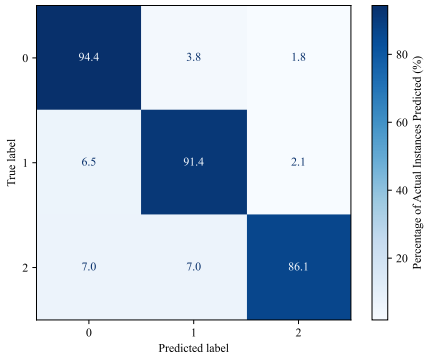


Figure 5.6: Confusion Matrix for k-NN (k=5)

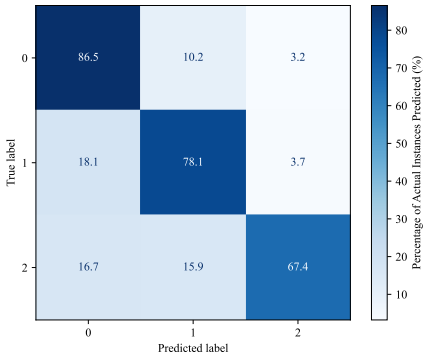


Figure 5.7: Confusion Matrix for k-NN (k=10)

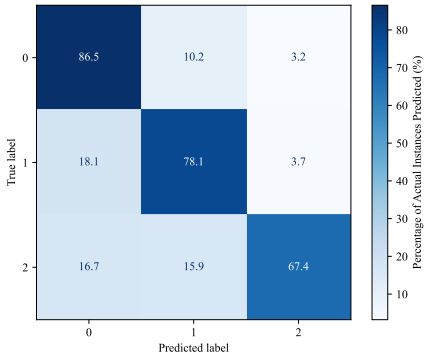


Figure 5.8: Confusion Matrix for k-NN (k=15)

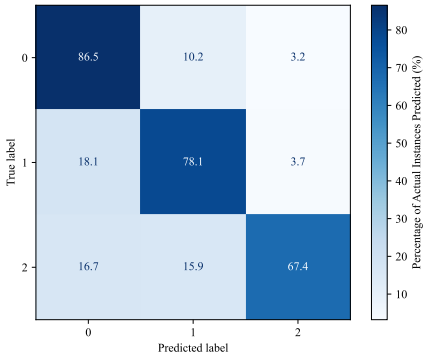


Figure 5.9: Confusion Matrix for k-NN (k=20)

The results, summarized in Table 5.5.

Model	Accuracy	F1 Score	Precision	Recall	AUC
Adaboost (Estimators=50)	0.647	0.644	0.653	0.647	0.814
Adaboost (Estimators=100)	0.702	0.701	0.708	0.702	0.829
Adaboost (Estimators=150)	0.720	0.718	0.724	0.720	0.844
Adaboost (Estimators=200)	0.732	0.731	0.736	0.732	0.849

Table 5.5: Mean Scores for AdaBoost Models with Different Numbers of Estimators

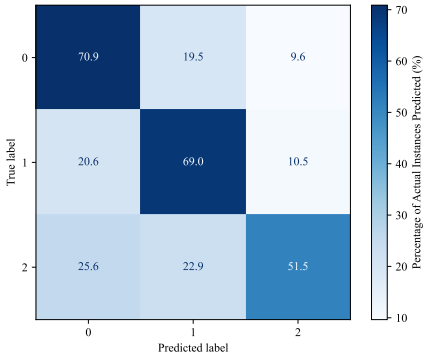


Figure 5.10: Confusion Matrix for AdaBoost (Estimators=50)

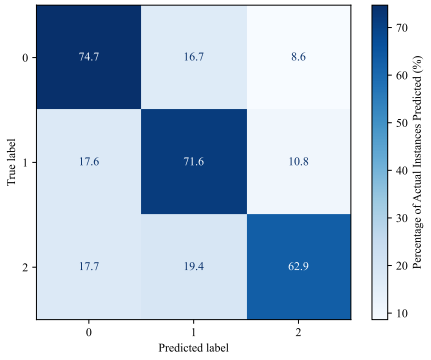


Figure 5.11: Confusion Matrix for AdaBoost (Estimators=100)

Neural Netwrok -MLP

Four MLP classifiers were trained with varying parameters to explore their impact on performance. The parameters used for each classifier are detailed in Table 5.6. MLP-1 utilized a hidden layer size of 200 neurons, while MLP-2 had a hidden layer size of 100 neurons. Both MLP-1 and MLP-2 employed the Rectified Linear Unit (ReLU) activation function and the Adam solver, with an alpha value of 0.0008 and a maximum of 400 iterations. Additionally, MLP-3 and MLP-4 utilized logistic activation functions.
The mean scores for the MLP classifiers are presented in Table 5.7. Notably, MLP-3 achieved the highest accuracy, F1 score, precision, recall, and Area Under the Curve (AUC) among all models, indicating superior performance.

Model	Hidden Layer Sizes	Activation	Solver	Alpha	Max Iter
MLP-1	(200,)	ReLU	adam	0.0008	400
MLP-2	(100,)	ReLU	adam	0.0008	400
MLP-3	(100,)	Logistic	adam	0.0008	400
MLP-4	(200,)	Logistic	adam	0.0008	400

Table 5.6: Parameters Used for MLP Classifiers

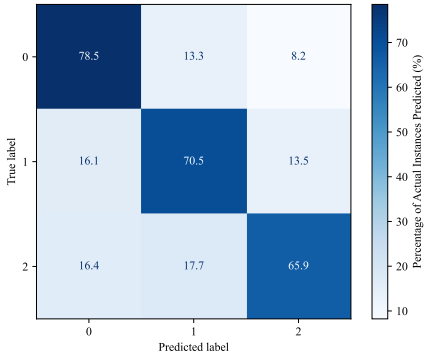


Figure 5.12: Confusion Matrix for AdaBoost (Estimators=150)

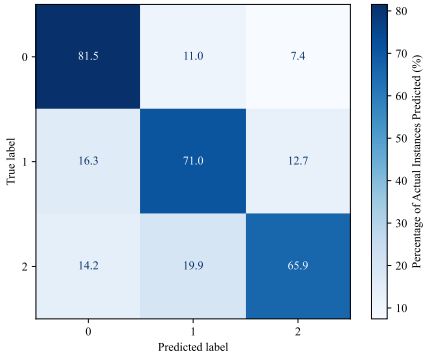


Figure 5.13: Confusion Matrix for AdaBoost (Estimators=200)

Model	Accuracy	F1 Score	Precision	Recall	AUC
MLP-1	0.894	0.894	0.895	0.894	0.972
MLP-2	0.889	0.889	0.891	0.889	0.969
MLP-3	**0.922**	**0.922**	**0.924**	**0.922**	**0.987**
MLP-4	0.909	0.909	0.910	0.909	0.982

Table 5.7: Mean Scores for MLP Classifiers

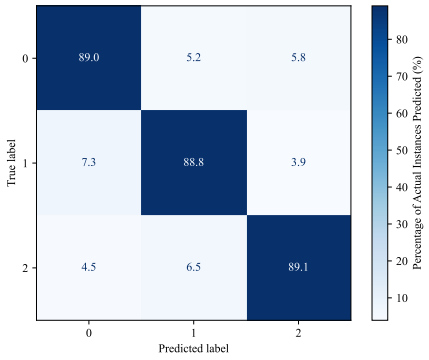


Figure 5.14: Confusion Matrix for MLP-1

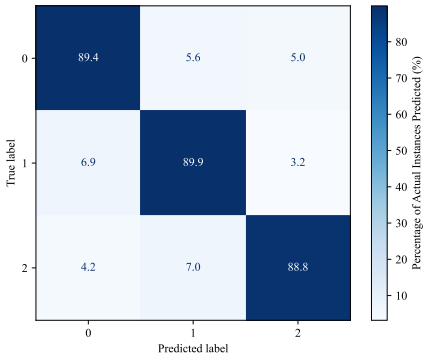


Figure 5.15: Confusion Matrix for MLP-2

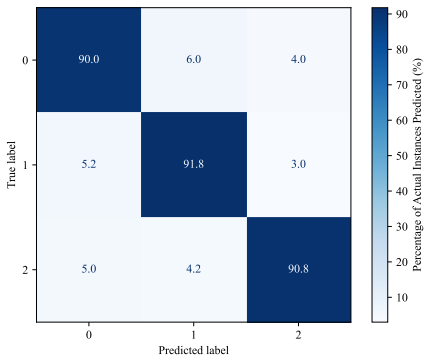


Figure 5.16: Confusion Matrix for MLP-3

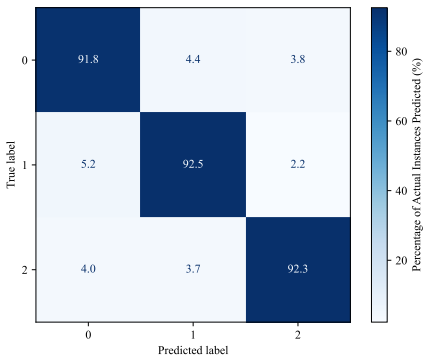


Figure 5.17: Confusion Matrix for MLP-4

6 Conclusion

6.1 Discussion

Discussion of the results compared to the objective— Limitation(sujective ground thruth)— - Future work(improvement in detection model, real time system schematic and breif, parameter tuning of the controller etc etc) and then conclusion

6.2 Analysis of ground truth

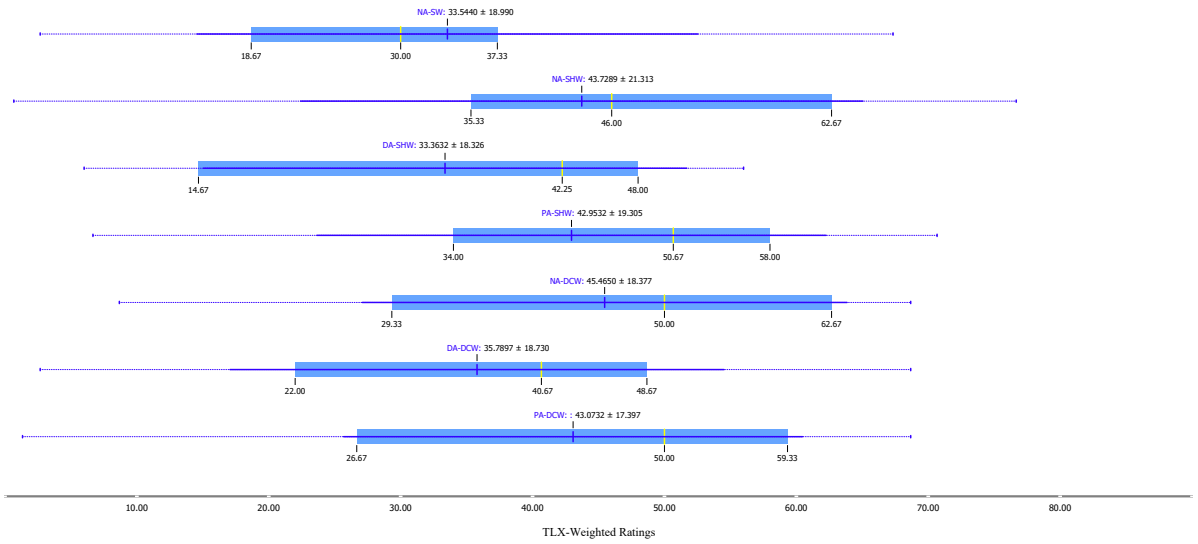


Figure 6.1

In our research on stress analysis, we primarily relied on subjective measures for labeling and establishing ground truth. While many previous studies have successfully employed this approach, it is important to acknowledge that subjective ratings can be influenced by a variety of factors. These influences could range from individual perception differences to contextual and environmental factors impacting a participant's response. Consequently, reducing the complex and multifaceted nature of stress into a three-class system based on subjective assessments might lead to oversimplifications of the nuanced nature of stress.

To bolster the reliability of our labeling process, we undertook an initiative to crossvalidate our labels with other physiological or behavioral stress indicators. This crossvalidation aimed to correlate our subjective stress labels with objective measures, such as mean skin conductance response meanscr and heart rate meanhr. It is crucial to note that this assumption that meanscr and meanhr directly indicate stress was specifically for the purpose of this crossvalidation exercise and not a general assumption throughout our project.
Given that physiological features like meanscr and meanhr are integral inputs of our stress detection model, they could not be directly used for labeling the stress categories. This constraint necessitated our reliance on subjective measures for labeling. However, by attempting to correlate these objective physiological indicators with our subjective labels, we sought to add a layer of validation to our approach.
Before delving into our results, it is imperative to discuss some critical caveats about our methodology. Our reliance on subjective ratings, though methodologically sound in many aspects, does not entirely capture the reliable ground truth of stress. The subjective nature of these ratings means they are susceptible to various influencing factors, which might not always align perfectly with physiological manifestations of stress. This acknowledgment is crucial for a comprehensive understanding of our findings and their interpretation.
Our approach, while aligned with standard practices in stress research, does acknowledge the limitations inherent in using subjective measures for stress categorization. By attempting cross-validation with physiological data, we endeavored to enhance the robustness of our methodology. Nevertheless, the complexities and intricacies of stress as a psychological and physiological phenomenon demand careful consideration and interpretation of our results.

6.3 Future work

Bibliography

Aigrain, J. (2016): "Multimodal detection of stress : evaluation of the impact of several assessment strategies. (Détection multimodale du stress : évaluation de l'impact de plusieurs stratégies de mesure)". In: URL: https://api.semanticscholar.org/CorpusID:3066188.

Aigrain, J., M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani (2018): "Multimodal Stress Detection from Multiple Assessments". In: *IEEE Transactions on Affective Computing* 9.4, pp. 491–506.

Alexander, D. M., C. Trengove, P. Johnston, T. Cooper, J. August, and E. Gordon (2005): "Separating individual skin conductance responses in a short interstimulus-interval paradigm". In: *Journal of neuroscience methods* 146.1, pp. 116–123.

Alshamrani, M. (2021): "An Advanced Stress Detection Approach based on Processing Data from Wearable Wrist Devices". In: *International Journal of Advanced Computer Science and Applications* 12.7. URL: http://dx.doi.org/10.14569/IJACSA.2021.0120745.

Arsalan, A., M. Majid, I. F. Nizami, W. Manzoor, S. M. Anwar, and J. Ryu (June 2023): *Human Stress Assessment: A Comprehensive Review of Methods Using Wearable Sensors and Non-wearable Techniques*. en. arXiv:2202.03033 [cs]. URL: http://arxiv.org/abs/2202.03033 (visited on 02/02/2024).

Badillo, S., B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, and J. D. Zhang (Apr. 2020): "An introduction to machine learning". en. In: *Clin. Pharmacol. Ther.* 107.4, pp. 871–885.

Bakhsh, A., G. F. J. Martin, C. D. Bicknell, C. Pettengell, and C. Riga (May 2019): "An evaluation of the impact of high-fidelity endovascular simulation on surgeon stress and technical performance". en. In: *J. Surg. Educ.* 76.3, pp. 864–871.

Bhushan, U. and S. Maji (2023): "Prediction and Analysis of Stress Using Machine Learning: A Review". In: *Proceedings of Third Doctoral Symposium on Computational Intelligence*, pp. 419–432.

BIOBSS (n.d.): *BIOBSS Python package*. Available under: https://biobss.readthedocs.io/en/latest/index.html.

Bradley, M. M. and P. J. Lang (1994): "Measuring emotion: The self-assessment manikin and the semantic differential". In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1, pp. 49–59. URL: https://www.sciencedirect.com/science/article/pii/0005791694900639.

Brantley, P. J., C. D. Waggoner, G. N. Jones, and N. B. Rappaport (Feb. 1987): "A Daily Stress Inventory: development, reliability, and validity". en. In: *J. Behav. Med.* 10.1, pp. 61–74.

Cohen, S., T. Kamarck, and R. Mermelstein (1983): "A Global Measure of Perceived Stress". In: *Journal of Health and Social Behavior* 24.4, pp. 385–396. URL: http://www.jstor.org/stable/2136404 (visited on 02/11/2024).

Dawson, M. E., A. M. Schell, and D. L. Filion (2007): "The electrodermal system." In: *Handbook of psychophysiology, 3rd ed.* Pp. 159–181.

Empatica (n.d.[a]): *Empatica E4*. Available under: https://e4.empatica.com/e4-wristband.

Empatica (n.d.[b]): *Empatica E4 specs*. Available under: https://box.empatica.com/documentation/20141119_E4_TechSpecs.pdf.

Empatica (n.d.[c]): *Utilizing the PPG/BVP signal*. Available under: https://support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal.

Favre-Félix, J., M. Dziadzko, C. Bauer, A. Duclos, J.-J. Lehot, T. Rimmelé, and M. Lilot (Aug. 2022): "High-fidelity simulation to assess Task Load Index and performance: A prospective observational study". en. In: *Turk. J. Anaesthesiol. Reanim.* 50.4, pp. 282–287.

Garbarino, M., M. Lai, D. Bender, R. Picard, and S. Tognetti (Jan. 2015): "Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition". In: pp. 39–42.

Gedam, S. and S. Paul (2021): "A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques". en. In: *IEEE Access* 9, pp. 84045–84066. URL: https://ieeexplore.ieee.org/document/9445082/ (visited on 02/02/2024).

Gellman, M. D. (2020): "Behavioral medicine". In: *Encyclopedia of behavioral medicine*, pp. 223–226.

Giakoumis, D., A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva (Sept. 2012): "Using Activity-Related Behavioural Features towards More Effective Automatic Stress Detection". In: *PLOS ONE* 7.9, pp. 1–16. URL: https://doi.org/10.1371/journal.pone.0043571.

Giannakakis, G., D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis (Jan. 2022): "Review on Psychological Stress Detection Using Biosignals". In: *IEEE Transactions on Affective Computing* 13.1. Conference Name: IEEE Transactions on Affective Computing, pp. 440–460. URL: https://ieeexplore.ieee.org/document/8758154 (visited on 01/08/2024).

Greco, A., G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi (2016): "cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing". In: *IEEE Transactions on Biomedical Engineering* 63.4, pp. 797–804.

Grier, R. A. (Sept. 2015): "How high is high? A meta-analysis of NASA-TLX global workload scores". en. In: *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 59.1, pp. 1727–1731.

Harrigan, J. A. (1985): "Self-touching as an indicator of underlying affect and language processes". In: *Social Science and Medicine* 20.11, pp. 1161–1168. URL: https://www.sciencedirect.com/science/article/pii/0277953685901935.

Hart, S. G. and L. E. Staveland (1988): "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Human Mental Workload*. Vol. 52. Advances in Psychology, pp. 139–183. URL: https://www.sciencedirect.com/science/article/pii/S0166411508623869.

Hernando-Gallego, F., D. Luengo, and A. Artés-Rodríguez (2017): "Feature extraction of galvanic skin responses by nonnegative sparse deconvolution". In: *IEEE journal of biomedical and health informatics* 22.5, pp. 1385–1394.

imotions (n.d.): *EDA Example signals*. Available under: https://imotions.com/blog/learning/research-fundamentals/eda/.

Jin, C. W., A. Osotsi, and Z. Oravecz (Dec. 2020): "Predicting Stress in Teens from Wearable Device Data Using Machine Learning Methods". In: URL: http://dx.doi.org/10.1101/2020.11.26.20223784.

Kaduk, S. I., A. P. J. Roberts, and N. A. Stanton (Jan. 2021): "Driving performance, sleepiness, fatigue, and mental workload throughout the time course of semi-automated driving—Experimental data from the driving simulator". en. In: *Hum. Factors Ergon. Manuf.* 31.1, pp. 143–154.

Koverola, M., A. Kunnari, J. Sundvall, and M. Laakasuo (June 2022): "General Attitudes Towards Robots Scale (GAToRS): A New Instrument for Social Surveys". In: *International Journal of Social Robotics* 14, pp. 1–23.

Kraaij, W., S. Koldijk, and M. Sappelli (2014): *The SWELL knowledge work dataset for stress and user modeling research.*

Krämer, M., C. Rösmann, F. Hoffmann, and T. Bertram (2020): "Model predictive control of a collaborative manipulator considering dynamic obstacles". In: *Optimal Control Applications and Methods* 41.4, pp. 1211–1232. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/oca.2599. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/oca.2599.

Kyriakou, K., B. Resch, G. Sagl, A. Petutschnig, C. Werner, D. Niederseer, M. Liedlgruber, F. Wilhelm, T. Osborne, and J. Pykett (Sept. 2019): "Detecting moments of stress from measurements of wearable physiological sensors". en. In: *Sensors (Basel)* 19.17, p. 3805.

Lagomarsino, M., M. Lorenzini, E. De Momi, and A. Ajoudani (2022): "An Online Framework for Cognitive Load Assessment in Industrial Tasks". In: *Robotics and Computer-Integrated Manufacturing* 78, p. 102380. URL: https://www.sciencedirect.com/science/article/pii/S0736584522000679.

Larsen, E., S. Gottschalk, M. Lin, and D. Manocha (Dec. 2000): "Fast Proximity Queries with Swept Sphere Volumes". In.

Lasota, P. A. and J. A. Shah (2015): "Analyzing the Effects of Human-Aware Motion Planning on Close-Proximity Human–Robot Collaboration". In: *Human Factors* 57.1. PMID: 25790568, pp. 21–33.

Lee, M.-h., G. Yang, H.-K. Lee, and S. Bang (2004): "Development stress monitoring system based on personal digital assistant (PDA)". In: *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* Vol. 1. IEEE, pp. 2364–2367.

Machado-Moreira, C., J. Caldwell Odgers, I. Mekjavic, and N. Taylor (Dec. 2008): "Sweat Secretion from Palmar and Dorsal Surfaces of the Hands During Passive and Active Heating". In: *Aviation, space, and environmental medicine* 79, pp. 1034–40.

Nahavandi, S. (2019): "Industry 5.0—A Human-Centric Solution". In: *Sustainability* 11.16. URL: https://www.mdpi.com/2071-1050/11/16/4371.

Nguyen, T. and Y. Zeng (Oct. 2017): "Effects of stress and effort on self-rated reports in experimental study of design activities". In: *Journal of Intelligent Manufacturing* 28.

Nikula, R. (1991): "Psychological Correlates of Nonspecific Skin Conductance Responses". In: *Psychophysiology* 28.1, pp. 86–90. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1991.tb03392.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1991.tb03392.x.

OptiTrack (n.d.): *Motive Skeleton Tracking.* Available under: https://docs.optitrack.com/motive/skeleton-tracking.

Pandian, V. and S. Suleri (Jan. 2020): *NASA-TLX Web App: An Online Tool to Analyse Subjective Workload.*

Pereira, R. and N. dos Santos (2023): "Reflections on a New Paradigmatic Approach for the Industry: A Scoping Review on Industry 5.0". In: *Logistics* 7.3. URL: https://www.mdpi.com/2305-6290/7/3/43.

Renz, H., M. Krämer, and T. Bertram (2023a): "Comparing Human Motion Forecasts in Moving Horizon Trajectory Planning of Collaborative Robots". In: *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1–6.

Renz, H., M. Krämer, and T. Bertram (2023b): "Uncertainty Estimation for Predictive Collision Avoidance in Human-Robot Collaboration". In: *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1–6.

Roberts, V. (1982): "Photoplethysmography- fundamental aspects of the optical properties of blood in motion". In: *Transactions of the Institute of Measurement and Control* 4.2, pp. 101–106. eprint: https://doi.org/10.1177/014233128200400205. URL: https://doi.org/10.1177/014233128200400205.

Sauppé, A. and B. Mutlu (2015): "The Social Impact of a Robot Co-Worker in Industrial Settings". In: URL: https://doi.org/10.1145/2702123.2702181.

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

Schulz, P. and W. Schlotz (Jan. 1999): "The Trier Inventory for the Assessment of Chronic Stress (TICS): Scale construction, statistical testing, and validation of the scale work overload". In: *Diagnostica* 45, pp. 8–19.

Setz, C., B. Arnrich, J. Schumm, R. Marca, G. Tröster, and U. Ehlert (Jan. 2010): "Discriminating stress from cognitive load using a wearable EDA device." In: *IEEE Transactions on Information Technology in Biomedicine* 14, pp. 410–417.

Sharma, N. and T. Gedeon (Dec. 2012): "Objective measures, sensors and computational techniques for stress recognition and classification: a survey". en. In: *Comput. Methods Programs Biomed.* 108.3, pp. 1287–1301.

Shukla, J., M. Barreda-Ángeles, J. Oliver, G. C. Nandi, and D. Puig (2021): "Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity". In: *IEEE Transactions on Affective Computing* 12.4, pp. 857–869.

Siirtola, P. and J. Röning (Aug. 2020): "Comparison of regression and classification models for user-independent and personal stress detection". en. In: *Sensors (Basel)* 20.16, p. 4402.

Smets, E., E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes, I. Van Diest, and C. Van Hoof (Dec. 2018): "Large-scale wearable data reveal digital phenotypes for daily-life stress detection". en. In: *NPJ Digit. Med.* 1.1, p. 67.

Stroop, J. R. (Dec. 1935): "Studies of interference in serial verbal reactions". en. In: *J. Exp. Psychol.* 18.6, pp. 643–662.

Suboh, M. Z., R. Jaafar, N. A. Nayan, N. H. Harun, and M. S. F. Mohamad (2022): "Analysis on Four Derivative Waveforms of Photoplethysmogram (PPG) for Fiducial Point Detection". In: *Frontiers in Public Health* 10. URL: https://www.frontiersin.org/articles/10.3389/fpubh.2022.920946.

Vos, G., K. Trinh, Z. Sarnyai, and M. Rahimi Azghadi (May 2023): "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review". en. In: *International Journal of Medical Informatics* 173, p. 105026. URL: https://linkinghub.elsevier.com/retrieve/pii/S1386505623000436 (visited on 01/10/2024).

Zakeri, Z., A. Arif, A. Omurtag, P. Breedon, and A. Khalid (Oct. 2023): "Multimodal assessment of cognitive workload using neural, subjective and behavioural measures in smart factory settings".

Zhang, Q., L.-G. Lindberg, R. Kadefors, and S. Jorma (May 2001): "A non-invasive measure of changes in blood flow in the human anterior tibial muscle". In: *Arbeitsphysiologie* 84, pp. 448–452.

7
Appendix

Das ist der Anhang (siehe Abschnitt **??**) / This is the appendix (see section **??**)

7.1 Usage of generative AI - Affidavit

☐ not at all

☒ for correcting, optimizing, or restructuring the entire work (This eliminates the need for explicit marking of individual passages or sections, as this type of usage refers to the entire written work. Explicit marking in the text is not necessary, as this serves as the global indication.)

☐ Code optimization: Optimization or restructuring of software function

☐ Code generation: Creating entire software functions from a detailed functional description.

☐ Substance generation in code: Generating entire software source code

☐ Media optimization: Correction, optimization, or restructuring of entire passages

☐ Media generation: Creating entire passages from given content.

☐ Substance generation in media: Generating entire sections

☐ More, namely:

I assure that I have provided all usages completely. Missing or incorrect information may be considered an attempt to deceive.

place, date

Jane Doe