Universität
Augsburg
University

In the Face and Heart of Data Scarcity in Industry 5.0: Exploring Applicability of Facial and Physiological AI Models for Operator Well-Being in Human-Robot Collaboration

DISSERTATION
zur Erlangung des Doktorgrades (Dr. rer. nat.)
an der
Fakultät für Angewandte Informatik
der Universität Augsburg

vorgelegt von

Pooja Prajod

2024

Erstgutachterin	Prof. Dr. Elisabeth André
Zweitgutachter	Prof. Dr. Johannes Schilp
Tag der mündlichen Prüfung	21.11.2024

Acknowledgements

This doctoral research journey has been a transformative experience, with many ups and some downs. I am deeply grateful to many individuals whose support and contributions have made this thesis possible.

First and foremost, I express my sincere gratitude to my supervisor, Prof. Dr. Elisabeth André, for her support and guidance throughout the course of my research. She has been a constant source of inspiration, which fueled both my personal and academic development. I would also like to thank Prof. Dr. Johannes Schilp for reviewing this thesis and providing his valuable feedback.

I am indebted to all my co-authors and colleagues from the Chair of Human-Centered Artificial Intelligence and the MindBot project. This thesis would not have been possible without their contributions. Special thanks to Matteo Lavit Nicora, Marta Mondellini, and Matteo Malosio from STIIMA CNR (National Research Council of Italy) for being excellent collaborators and integral part of many studies presented in this thesis. I would also like to extend my thanks to Kathrin Janowski, Hannes Ritschel, Katharina Weitz, Stina Klein, and Bhargavi Mahesh for their immense support and helpful discussions throughout these years.

Finally, I would like to express my deepest gratitude to my family and friends for their unconditional love and encouragement. I am eternally grateful to my husband, Sanchar Sharma, for his constant encouragement and for bringing out the best in me.

Pooja Prajod

Abstract

Over the past decade, research has focused on integrating collaborative robots, or cobots, into assembly lines. The envisioned future industrial workplaces involve close collaboration between human workers and cobots. With the advent of Industry 5.0, human-centered approaches to facilitate human-robot collaboration (HRC) have gained significant traction. These approaches go beyond ensuring physical safety, emphasizing the mental health and well-being of industrial workers. To achieve this goal, cobots have to be equipped with capabilities to detect real-time worker states. Despite various investigations into user states related to well-being in different domains, the manifestations of these states in industrial settings are relatively unexplored. Hence, a critical gap exists in our understanding of whether machine learning models developed for other contexts are applicable to industrial HRC.

Many aspects of existing datasets pose challenges to the applicability of the machine learning models in industrial settings. On the one hand, most datasets for well-being-related states (e.g., pain, distraction) are typically small and lack variation in recording conditions, raising concerns about whether models trained on these datasets learn generic or dataset-specific features. On the other hand, although states like stress are well-researched, there are limited public datasets involving HRC tasks. This limitation is exacerbated by the lack of long-term studies involving industrial HRC tasks, limiting our understanding of worker states (e.g., boredom, flow) that emerge over a long period of familiar and repetitive tasks. These limitations of existing datasets form the motivation for the works presented in this thesis.

This thesis explores applicability through multiple lenses: transferability (leveraging features from a related task), generalizability (ensuring models perform well on multiple datasets), replicability (testing approaches on various datasets and recording conditions), reproducibility (recreating industrial HRC experiences), and versatility (utilizing features/-models for multiple tasks). The investigations of this thesis are presented in two parts. The first part addresses transferability, generalizability, and replicability by utilizing transfer learning techniques to train various models and assess them using explainable AI methods and cross-dataset evaluations. The second part addresses reproducibility and versatility by analyzing user studies in simulated industrial HRC scenarios with durations ranging from half an hour to several days. The results of this thesis not only demonstrate approaches to develop models applicable to industrial HRC settings but also identify potential avenues for improvement. These findings form the foundations for developing models that enhance human-robot collaboration in industrial environments by focusing on both efficiency and worker well-being.

Acronyms

ANN	Artificial Neural Network
ANS	Autonomic Nervous System
ASD	Autism Spectrum Disorder
AU	Action Unit
BVP	Blood Volume Pulse
CNN	Convolutional Neural Network
ECG	ElectroCardioGram
EDA	ElectroDermal Activity
ELoC	Experiential Locus of Control
HRC	Human-Robot Collaboration
HRV	Heart Rate Variability
LOSO	Leave One Subject Out
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory
RFC	Random Forest Classifier
SAM	Self-Assessment Manikin
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TLX	Task Load IndeX
TSST	Trier Social Stress Test
XAI	eXplainable Artificial Intelligence

Contents

Part I

Foundation

Chapter 1

Introduction

1.1 A Brief History of Industrial Revolutions

Throughout history, distinct industrial revolutions have emerged, each bringing about drastic improvements in workflow and productivity [Xu et al., 2018; Popkova et al., 2019; Sharma and Singh, 2020]. The first industrial revolution, or Industry 1.0, began in the late 18^{th} century and was characterized by the transition from manual production methods to machine-based manufacturing. The invention of steam engines powered the rapid mechanization of factories. The second industrial revolution, or Industry 2.0, occurred in the late 19^{th} and early 20^{th} centuries. This period saw the widespread use of electricity and mass production techniques like the assembly line. The third industrial revolution, or Industry 3.0, began in the late 20^{th} century and was driven by the development of digital technologies. This revolution saw computer-aided automation become an integral part of the industrial setting. The fourth industrial revolution, or Industry 4.0, began around 2010 and led to the adoption of advanced technologies such as the Internet of Things, Machine Learning, and Cloud Computing. This revolution focused on end-to-end digitalization, paving the way for “smart factories". One of the primary technologies of Industry 4.0 is Cyber-Physical Systems called Collaborative Robots, or Cobots. These are robotic arms that can work safely alongside human workers.

Currently, factories are transitioning towards a human-centric paradigm as part of the fifth industrial revolution or Industry 5.0 [Sharma and Singh, 2020; Xu et al., 2021; Adel, 2022; Rožanec et al., 2023]. Unlike previous industrial revolutions, where each era largely superseded the last, Industry 5.0 coexists with Industry 4.0. Industry 5.0 can be seen as an extension of Industry 4.0, leveraging similar technologies but focusing on human-centered design and collaboration. This new revolution emphasizes close collaboration between human workers and cobots, leveraging each of their individual strengths. Imagine a cobot and a worker sharing a workspace, simultaneously working on an object – this exemplifies the collaborative workflow of Industry 5.0.

In the famous words of the Greek philosopher Aristotle - human beings are by nature social animals. Humans thrive on interaction and can struggle in isolated environments. Unfortunately, many factory work cells, especially small- and medium-scale industries, involve a single worker collaborating with the robot, limiting human-to-human interactions [Hovens, 2020; Mühlemeyer, 2020; Osika, 2023]. This reduction in social interactions can lead to emotional loneliness, which might negatively impact employee performance [Akçit and Barutçu, 2017]. Moreover, the reduced interactions with human colleagues also hinder skill development due to knowledge sharing, which slows down the technology acceptance of cobots [Welfare et al., 2019; Meissner et al., 2020].

Beyond social isolation, other factors can also impact the workers' well-being in Human-Robot Collaboration (HRC) settings. These factors include the speed of the cobot and the potential for an imbalanced workload, which can lead to stress, boredom, and frustration for the human worker [Welfare et al., 2019; Meissner et al., 2020; Storm et al., 2022]. These negative experiences can ultimately hinder the success of HRC implementations.

Hence, a foundational principle of Industry 5.0 is that optimal human-robot communication and worker well-being lead to improved productivity [Grosse et al., 2023; Loizaga et al., 2023]. When workers feel supported by prioritizing their mental health, they are more likely to be engaged, efficient, and less stressed. So, it is crucial to equip cobots with capabilities to recognize the workers' experiences, transforming them into Cyber-Physical-Cognitive systems [Adel, 2022; Rožanec et al., 2023; Khan et al., 2023]. By understanding the worker's affective and cognitive states, a cobot can adapt its behavior to be more supportive and improve the working environment.

1.2 Workers' Mental Health and Well-Being

Worker safety has always been paramount when introducing cobots into factories [Khalid et al., 2017; Robla-Gómez et al., 2017; Bragança et al., 2019; Arents et al., 2021; Coronado et al., 2022; Gladysz et al., 2023]. Initially, cobots were separated from workers by physical barriers to ensure the workers' safety. However, Industry 5.0 emphasizes close collaboration between humans and cobots. This shift resulted in a renewed focus on ensuring worker safety through cobot behavior. For instance, cobots are now equipped with strategies to prevent collisions with workers, minimizing the risk of physical harm. These strategies not only ensure worker safety but also foster trust in cobot technology, ultimately leading to acceptance of cobots among workers and its adoption in factories [Lu et al., 2022b; Baltrusch et al., 2022; Faccio et al., 2023].

However, human-centricity in Industry 5.0 is not limited to the physical safety of the workers. As mentioned before, the introduction of cobots can reduce face-to-face interactions between workers, potentially leading to feelings of loneliness and a lack of social support, both of which are critical for mental well-being [Rook, 1985; Leigh-Hunt et al., 2017; Saltzman et al., 2020; Emerson et al., 2021]. Ideally, cobots would evolve to serve the role of a supportive colleague, capable of social interaction with the worker [Storm et al., 2022; Osika, 2023; Sharma et al., 2024]. This expanded role goes beyond simply incorporating a human-like appearance for the cobot. That is, cobots should be capable of perceiving and responding to social cues from the worker [Sauppé and Mutlu, 2015; Nicora et al., 2021; Dwyer et al., 2021; Baltrusch et al., 2022]. For example, a cobot that detects stress in a worker could adapt its movement speed or offer assistance, fostering a more positive and engaging work environment. Such adaptations would need the cobots to leverage machine learning models to interpret non-verbal cues such as facial expressions, body language, and physiological signals (e.g., heart rate, respiration, etc.) [Papetti et al., 2020; Gervasi et al., 2023; Loizaga et al., 2023; Gladysz et al., 2023].

Stress is one of the most researched experiences that negatively impacts the mental well-being of workers [Tran et al., 2022; Lu et al., 2022a; Coronado et al., 2022; Blandino et al., 2023; Faccio et al., 2023; Loizaga et al., 2023; Adattil et al., 2024]. There are many sources of stress in an industrial setting. Industrial environments can be inherently stressful due to noisy manufacturing processes or higher temperatures. Chronic exposure to these conditions can contribute to stress and fatigue [Battini et al., 2022; Adattil et al., 2024]. The rapid adoption of cobots can also be a significant source of stress for workers. They may fear that cobots will render their skills obsolete, leading to the loss of their jobs [El Makrini et al., 2018; Fedorova et al., 2022; Liao et al., 2023]. Moreover, the increasing complexity of cobot technology can be a source of "techno-stress" [Brod, 1984; Shu et al., 2011] as workers may struggle to understand and adapt to unfamiliar systems [Wurhofer et al., 2015; Zambon, 2022; Cunha et al., 2022].

In addition to these long-term stressors, the cobots' behaviors can contribute to worker stress. Cobots operating close to the worker or moving at high speeds can significantly increase a worker's cognitive load [Kato et al., 2010]. The worker may feel the need to constantly monitor and adapt to the robot's actions, leading to mental fatigue and stress. Even cobot adaptations designed for worker safety, such as changes in speed or trajectory, might be misinterpreted as unpredictable behavior and lead to additional stress.

Researchers have been exploring ways for prophylactic facilitation of workers' well-being. One area of focus is detecting physical fatigue and similar ergonomic factors [Gualtieri et al., 2021; Coronado et al., 2022; Faccio et al., 2023; Loizaga et al., 2023]. Cobots equipped with sensors to detect fatigue patterns of workers (e.g., changes in posture, decreased movement) could dynamically adjust their behavior to support the worker. For instance, a cobot could take on a higher share of the workload or suggest breaks at appropriate times.

An emerging topic in prophylactic cobot adaptations relies on pain detection [Giallanza et al., 2024]. The typical concept involves the cobot adapting to the onset of pain in workers and is particularly critical for ensuring safety. Additionally, Meissner et al. [2020] reported that industrial workers develop muscle and joint pain due to repetitive tasks, which can be alleviated by cobots sharing the workload. Moreover, social support is a crucial in mitigating work-related pain [Baek et al., 2018; Hoogendoorn et al., 2000]. So, detecting pain is a prerequisite for designing cobots that can provide appropriate social support to workers.

Another well-researched human factor involves worker intent recognition [Lu et al., 2022b; Mukherjee et al., 2022; Rožanec et al., 2023]. Cobots that can infer a worker's intended actions can improve safety and productivity. For example, a cobot anticipating the worker's movement can adjust its trajectory to avoid collisions. Similarly, a cobot that understands which object a worker intends to use next could proactively prepare that object and streamline the workflow. So, intent recognition and subsequent cobot behavioral adaptations can contribute to reducing painful and stressful situations.

Researchers are also exploring the potential of worker attention/distraction and emotion recognition in the context of HRC [Toichoa Eyam et al., 2021; Mukherjee et al., 2022; Coronado et al., 2022; Loizaga et al., 2023]. Cobots that can understand a worker's level of focus or emotional state could tailor their interactions accordingly. For instance, the cobot that can detect boredom may offer to take up repetitive tasks so that the worker can engage in a more stimulating task. However, it's important to note that research on worker attention/distraction and emotion recognition in industrial settings is still in its early stages.

1.3 Need for Assessing Applicability

Based on the literature discussed in the previous section, this thesis initially focuses on three worker states relevant to industrial HRC settings: distraction, pain, and stress. The next steps involve identifying modalities for detecting these states and training machine learning models. While such models are promising for improving cobot's supportive behavior, the following challenges regarding the identified states and the industrial HRC context need to be addressed.

1.3.1 Lack of Public Datasets from Industrial HRC

As highlighted earlier, states like distraction and pain are not frequently researched in Industry 5.0. This results in a scarcity of publicly available datasets collected during industrial HRC scenarios. Even well-researched areas like stress detection often lack publicly available datasets from real factory workers or realistic simulations. Consequently, training models for these states often rely on public datasets from other contexts. Inevitably, this leads to potential discrepancies in the characteristics of the experienced state. For example, the methods used to elicit the state (sudden versus existing pain, cognitive versus social stress) or the intensity of the state (moderate versus high) may differ, leading to variations in participants' responses. These differences raise the question: **Can models trained on datasets from different contexts be effectively applied to industrial HRC settings? Or are these models specific to the training context?**

1.3.2 Small Size of Datasets

Public datasets for some states, like pain, are typically small due to ethical considerations involved in inducing such states. Additionally, subjecting participants to prolonged negative states raises health concerns. Another reason for the limited dataset size could be the genuine rarity of certain states or the fact that their manifestations occur only after extended periods. For instance, natural distractions due to monotony might emerge only after performing the task for an extended duration. Initially, the task might be engaging due to its novelty or the learning involved in getting familiar with it. Mind wandering and distractions occur after these factors diminish over time. While techniques exist to train well-performing models on small datasets, the limited variation within these datasets raises another question: **Are the features learned by the model generic and applicable to real-world scenarios, or are they specific to the limited data available?**

1.3.3 Limited Long-term Studies

The scarcity of public industrial HRC datasets is further compounded by the lack of long-term studies, even in lab settings mimicking industrial environments. While the survey papers on Industry 5.0 reveal some relevant worker states, most of the works they analyzed were limited in duration (typically less than one hour) and may not capture the full range of worker experiences. For example, tiredness or apathy often manifests only after extended periods. This limitation raises two key questions: **What other relevant worker states might manifest during long-term industrial HRC? Can machine learning models be applied to detect these newly identified states?**

1.4 Research Objectives

The core focus of this thesis is to develop and train machine learning models for real-time cobot behavior adaptations in industrial HRC settings, while simultaneously addressing the overarching challenges identified earlier. These challenges necessitate novel approaches to model development and assessment of applicability. The various facets of applicability, presented in the upcoming chapters, are discussed below.

1.4.1 Transferability of Learned Features

Deep learning models are currently a popular choice for various tasks due to their ability to achieve high prediction performances. However, training these models often requires substantial amounts of data. As identified earlier, some relevant states often have small datasets, which poses a significant challenge for deep learning model development.

Transfer learning is a technique that can be leveraged to address this challenge. It involves *applying* the knowledge learned by a model on a source task to train models for a related target task. The transfer learning process for worker state detection can be broken down into several steps:

- **Identify suitable modalities**: A wide range of modalities, such as facial images, physiological signals, and gaze data, have been explored in the literature for detecting various user states. However, not all modalities are equally suitable for industrial environments. For example, audio data may be impractical due to noise levels, and finger-based heart beat sensors might be disruptive during assembly tasks. Hence, identifying data modalities suitable for industrial HRC is an essential first step.

- **Identify source datasets**: Ideally, source datasets for transfer learning should: (a) be a large dataset and (b) share some similarities with the target dataset. For instance, a large dataset of labeled facial images could serve as a source for training a smile detection model on a smaller dataset. This step will explore various existing datasets that meet these criteria and are appropriate for the target task.

- **Train source models**: Once suitable source datasets are identified, deep learning models will be trained using these datasets. This step involves defining the deep learning network architecture and empirically setting the training parameters to optimize model performance on the source task.

- **Train and evaluate target models**: Finally, the target models (e.g., pain detection model) will be trained on the smaller target dataset, leveraging the knowledge learned from the source models. The performance of these target models will be evaluated and compared to existing state-of-the-art models.

1.4.2 Generalizability of Models

Generalizability is a critical aspect of machine learning model development. It refers to a model's ability to perform well on unseen data, indicating that it has learned generic features *applicable* beyond the training data. Evaluating a model's generalizability is essential to ensure its real-world *applicability* in industrial HRC settings.

Some studies evaluate generalization using leave-one-subject-out (LOSO) or hold-out participants. This evaluation method involves training the model on data from a subset of participants and evaluating its performance on the remaining unseen participants. While this approach provides a basic assessment of generalizability, it is important to note that the overall dataset characteristics (e.g., recording settings, elicitation method) remain consistent across all participants.

Cross-dataset evaluations is a more rigorous assessment of generalizability. In this method, the model is trained on one dataset and then evaluated using a different dataset. If the model performs well on this new dataset (with potentially different elicitation methods, participant demographics, etc.), it suggests that the model has learned generic features.

Visualizing the features or parts of the input data that are most influential in the model's prediction can be used to assess generalizability. These visualizations can be generated using eXplainable Artificial Intelligence (XAI) techniques. Analyzing these visualizations can provide insights whether the model is relying on generic features or overfitting to specific characteristics of the training data.

This thesis will explore the above methods to assess the generalizability of the trained models. The assessment process will involve the following steps:

- **LOSO or hold-out evaluations**: All the models presented in this thesis will be evaluated using this method.

- **Cross-dataset evaluations**: The models developed for distraction, pain, and stress will be assessed for their generalizability using this method. The states identified through long-term studies are not evaluated through this method due to the lack of suitable datasets.

- **Devise XAI-based approach**: While existing XAI approaches rely solely on manual inspection of visualizations, this thesis proposes a novel systematic method for quantitatively assessing the learned features of a model. This new XAI-based approach will be demonstrated in the context of the pain detection models.

- **Determine factors contributing to low generalizability**: Generalizability assessments typically focus on determining if a model has learned generic features can be deployed in other contexts. This thesis will take an additional step to identify and analyze the specific factors within the datasets that may have contributed to lower generalizability of models.

1.4.3 Replicability of Methods

Unlike the previously discussed aspects related to model characteristics, replicability focuses on the methods. It refers to the ability to *apply* the developed methods effectively to different datasets, ensuring they are not specific to a particular dataset or its characteristics (e.g., elicitation method, participant demographics, or behavioral patterns). Replicability is crucial in research as it allows other researchers to reproduce the findings and extend the knowledge base.

Replicability of a method is typically demonstrated by successfully *applying* it to multiple datasets and obtaining similar results. In the context of this thesis, replicability encompasses the transfer learning processes employed, the generalizability assessment approaches developed, and other methodological aspects. For example, the source datasets and transfer learning steps used for training a facial pain detection model can be leveraged for training pain detection models using other facial pain datasets, even those not used in this thesis. This ensures the broader *applicability* of the transfer learning methods beyond the specific datasets employed here.

To emphasize the importance of replicability, all the methods developed in this thesis were tested on at least two datasets. This approach strengthens the confidence that these methods can be effectively utilized in various research contexts.

1.4.4 Reproducing Long-term States in Industry-like HRC Settings

As highlighted earlier, the scarcity of long-term (lasting for many hours and spanning over multiple days) industrial HRC studies presents a significant challenge in understanding the full range of worker experiences in these settings. This thesis tackles this challenge by reproducing an industrial HRC task within a laboratory environment. Two key considerations motivated the selection of a lab setting over a real-world factory environment. Firstly, real-world data collection involving video recordings or in-depth analyses of worker performance and working styles raises ethical concerns. In a factory environment, the data will be limited to the notes taken during fly-on-the-wall observations, hindering a comprehensive analysis of worker experiences. Secondly, the lab setting enables balanced recruitment, ensuring gender diversity and the inclusion of participants with Autism Spectrum Disorder (ASD). By leveraging a more inclusive participant pool, this thesis will explore the extent of *applicability* of solutions while taking into account the needs of ASD operators.

This thesis employed the following steps to identify the relevant long-term worker experiences in industrial HRC settings:

- **Identify relevant literature**: A review of existing literature will be conducted to explore user experiences and the associated manifestations during HRC tasks. The literature review will be extended to include studies involving individuals with ASD to identify specific behavioral patterns that may be relevant in an industrial context. The insights gained from this literature review will inform the selection of analysis tools and the behavioral aspects targeted for further investigation.

- **Design an industry-like setting involving a collaborative task**: To study long-term worker experiences, a laboratory environment mimicking an industry-like work cell will be designed. This setup will feature a well-defined collaborative assembly task. In line with the close collaboration envisioned in Industry 5.0, the task will involve a human operator and cobot: (a) working towards a common goal, (b) sharing the workspace at the same time, and (c) jointly manipulating an object. The task will be repeated over multiple production cycles, simulating a work shift that spans several hours (around four hours) of collaborative work. The participants will perform these simulated work shifts daily for one week to emulate long-term worker experiences.
- **Analyze collaboration sessions**: Given the exploratory nature of this objective, the experiences and behavioral patterns of the operators will be analyzed using a combination of quantitative and qualitative methods. This analysis approach will facilitate the identification of relevant experiences and repetitive behavioral patterns associated with long-term industrial HRC.

1.4.5 Versatility of Models and Features

Here, versatility refers to the *application* of features or models developed for one prediction task to another concept. This is particularly valuable for developing computationally efficient systems, as it leverages existing knowledge for new prediction tasks. For example, consider a model trained to detect stress, a state characterized by high arousal and elevated heart rate. The features learned by this model can potentially be leveraged to identify other high-arousal states.

This thesis will explore the versatility of the developed models and features, particularly in predicting newly identified long-term worker experiences. Assessing the versatility of models relies on the following steps:

- **Identify literature for newly identified experiences**: The first step involves identifying relevant literature related to the long-term experiences identified earlier. This literature review will focus on understanding the physiological and behavioral characteristics associated with these experiences. Acknowledging that these long-term states are rarely studied in HRC contexts, the review will be extended to literature from other relevant domains.
- **Inducing targeted experiences and behavioral patterns**: Informed by the findings from the literature review, the previously developed industry-like HRC setup will be adapted to elicit the identified long-term states within relatively shorter sessions. This adaptation will involve modifying the task characteristics and robot behavior to facilitate the emergence of these states. During the adapted task sessions, relevant data will be collected from participants for analysis and development of models.

- **Analyze participants' responses**: The physiological and behavioral responses from participants will be analyzed to assess whether the models already trained for other states or extracted features can be leveraged for detecting the long-term states.
- **Demonstrate models**: The final step involves demonstrating the versatility of the models and features. This thesis will consider two approaches to accomplish this step. The first approach leverages existing knowledge by utilizing the feature extraction from previously developed models. The extracted features are used to train new models for detecting the long-term states. The second approach deploys existing models for real-time cobot adaptation to address long-term states, thus showcasing the existing models' broader applicability.

1.5 Thesis Overview

This thesis is organized into four parts, as illustrated in Figure 1.1. A brief overview of the contents of each part is listed below:

- **Part 1 - Foundation (Chapters 1 and 2)**: This part lays the groundwork for the research by providing essential background information and introducing key concepts. The current chapter serves as the introduction, outlining the challenges addressed in the thesis. Chapter 2 delves deeper, introducing psychological definitions for various operator states, the physiological and behavioral responses associated with these states, and fundamental machine learning concepts used in training prediction models.
- **Part 2 - Transferability, Generalizability, and Replicability (Chapters 3, 4, and 5)**: This part presents the research outcomes obtained while developing models for predicting specific worker states: Attention/Distraction (Chapter 3), Pain (Chapter 4), and Stress (Chapter 5). These chapters address the challenges associated with the limited availability of public industrial HRC datasets and the issue of small dataset sizes. The developed models, methodologies, and insights contribute significantly to the research objectives related to the transferability, generalizability, and replicability of the developed solutions.
- **Part 3 - Reproducing Industrial HRC and Versatility (Chapters 6, 7, and 8)**: This part addresses the challenge of limited research on long-term worker experiences within industrial HRC settings. In contrast to Part 2, this part focuses on inducing specific states and behavioral patterns solely through cobot behavior manipulations. Chapter 6 describes a long-term industrial HRC study conducted in a laboratory environment. This study aims to reproduce worker states that typically manifest over extended periods in real-world settings. Chapter 7 explores the elicitation of flow, boredom, and anxiety by varying the production rate of the cobot. It then presents the analysis of participants' responses during these states and the development of a machine-learning model to differentiate them. Chapter 8 investigates the potential of using participants' gaze cues to initiate collaborative activities within the HRC tasks.

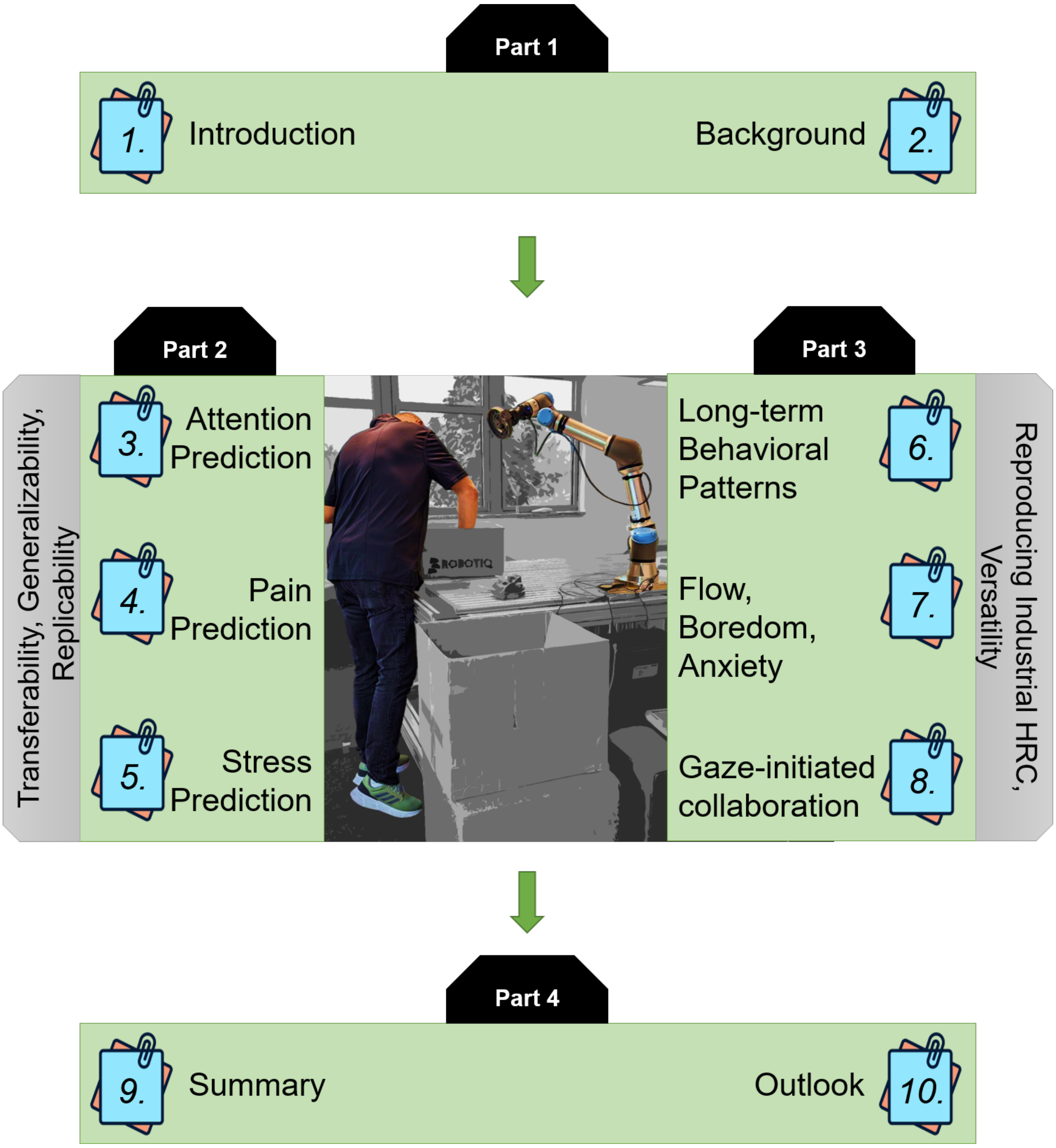


Figure 1.1: An overview of the parts and chapters of this thesis.

- **Part 4 - Conclusion (Chapters 9 and 10)**: The final part of the thesis consists of two chapters: Summary (Chapter 9) and Outlook (Chapter 10). Chapter 9 provides an overview of the key takeaways from Parts 2 and 3. It also highlights the contributions made by this thesis to the field of industrial HRC research. Chapter 10 outlines potential future directions to expand upon the presented research.

Chapter 2

Background

The research objectives of this thesis are centered around developing machine-learning models that are applicable to Human-Robot Collaboration (HRC) scenarios. This chapter establishes the essential background knowledge for the research presented in this thesis. It delves into three key areas: psychological concepts, measurable affective signals, and machine learning. This thesis utilizes behavioral and physiological signals that can serve as objective indicators of worker state. Behavioral signals often refer to various observable non-verbal communication cues, such as facial expressions, gaze, and body language [Arya et al., 2021; Lin and Li, 2023]. Physiological signals, on the other hand, are measurable biological signals produced by the body that provide insights into an individual's physical and mental condition [Arya et al., 2021; Lin and Li, 2023]. These biosignals such as electrocardiogram (ECG), blood volume pulse (BVP), and electrodermal activity (EDA) are often recorded using specialized sensors. The section on psychological concepts delves into various theories related to the worker states discussed later. Following this, the behavioral signals section examines the various behavioral patterns and modalities that can be used to infer relevant worker states. Similarly, the section on physiological signals discusses the various biosignals and the worker states that can be detected using them. Finally, the section on machine learning provides an overview of various model architectures and training techniques utilized in this thesis.

2.1 Psychological Concepts

This section outlines the psychological models and theories associated with some key worker states discussed in the subsequent chapters. It is important to note that a few states were excluded from this discussion due to the lack of well-established psychological theories associated with them.

2.1.1 Emotion

Human emotions play a critical role in the HRC experience. However, capturing and representing the complexities of human emotions presents a significant challenge for machine learning models. Psychology researchers have proposed two main approaches to address this challenge: discrete and continuous emotion models [Bota et al., 2019; Wang et al., 2022]. Discrete models categorize emotions into a finite set, while continuous models represent emotions as values within a multi-dimensional space.

Discrete Emotion Models

One of the first attempts to formulate a standard representation of emotions was by Ekman [1971], who proposed a set of six basic emotions: happy, sad, surprise, fear, anger, and disgust. However, later research supports the inclusion of contempt [Ekman and Heider, 1988; Paul Ekman Group, 2023]. Ekman's basic emotions are said to be instinctive, cross-cultural, and associated with well-established facial expressions [Gu et al., 2019; Wang et al., 2022]. Facial expressions associated with these emotions are visualized in Figure 2.1.

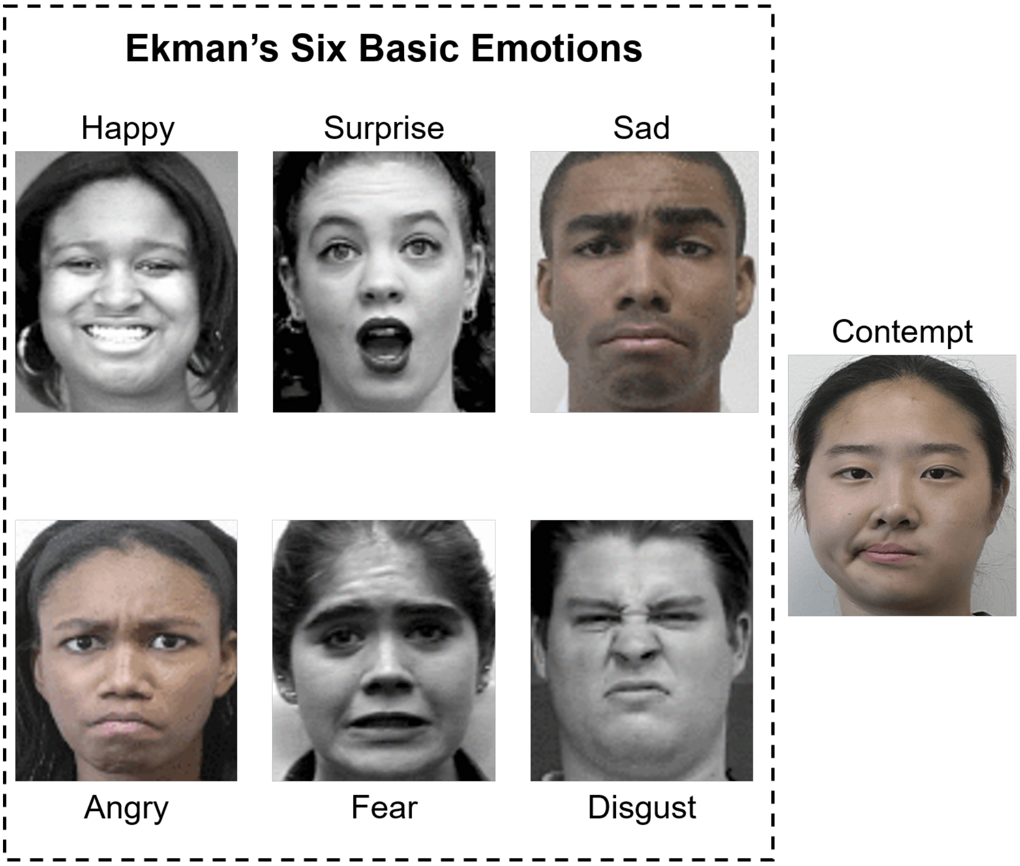


Figure 2.1: Images from CK+ [Lucey et al., 2010] facial expression dataset showing Ekman's six basic emotions and contempt (considered as a basic emotion later). Reused with permission, the copyright belongs to the publisher (Copyright©2010 IEEE)

Another popular discrete emotion model is Plutchik's emotion wheel model [Plutchik, 2003]. This model has a set of eight basic emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. A combination of these basic emotions forms more complex emotions. For example, joy and anticipation combine to form optimism. The model also incorporates intensity or levels of basic emotion, with more intense versions of the basic emotions closer to the center and less intense versions on the outer ring of the wheel. For example, rage is a more intense version of anger, whereas annoyance is a less intense version. Figure 2.2 depicts the wheel containing basic emotions, varying intensities, and complex emotions. Interestingly, a study by Yamashita and Kudoh [2022] comparing the two discrete emotion models suggests that Ekman's model aligns better with human intuition of basic emotions.

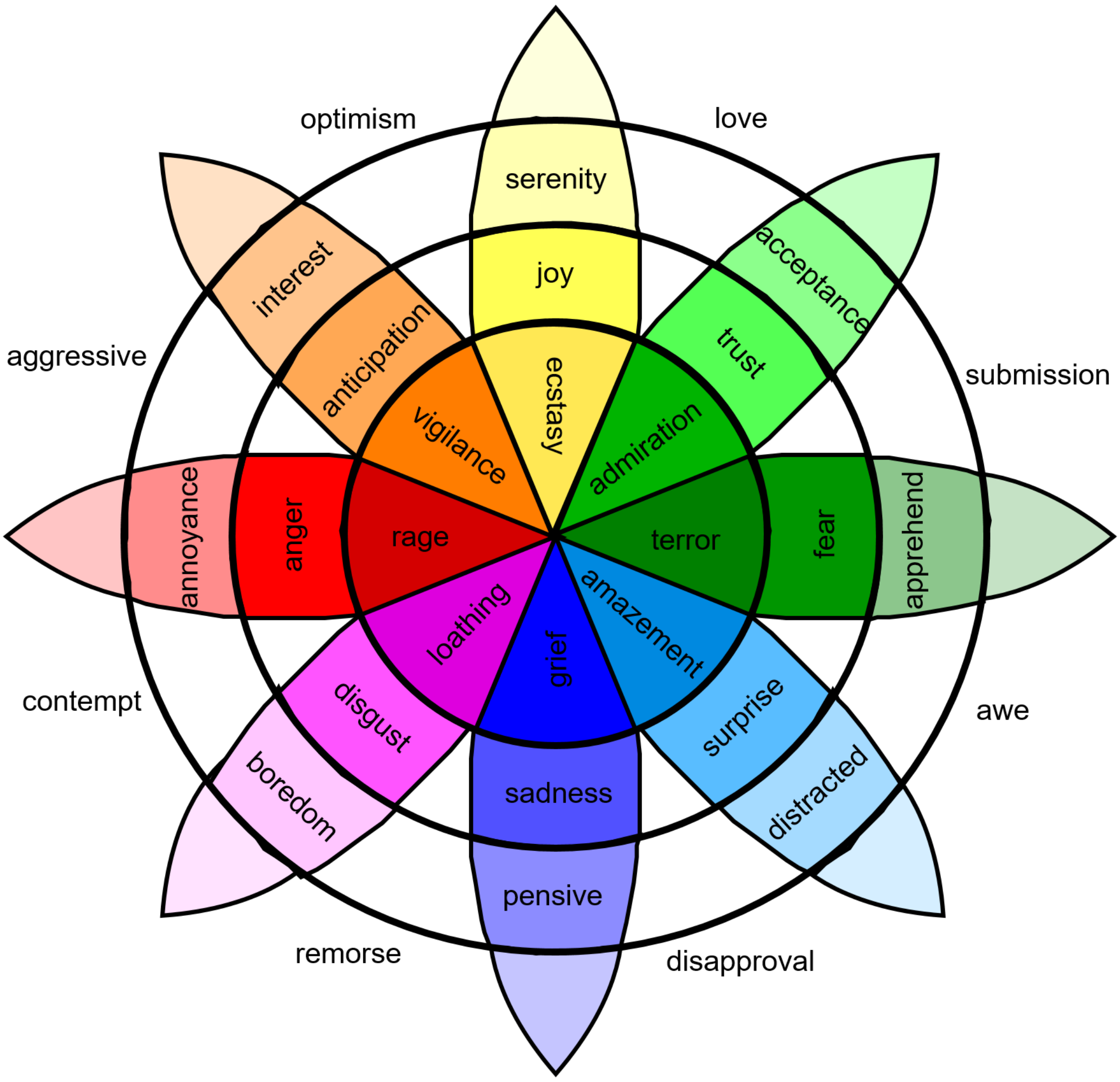


Figure 2.2: An illustration of Plutchik's wheel of emotions.

Continuous Emotion Models

A widely used dimensional emotion model is Russell's circumplex model [Russell, 1980]. As illustrated in Figure 2.3, this model represents emotions along two dimensions: valence and arousal. Valence (or pleasure) refers to the pleasantness of an emotion, with negative emotions like sadness and anger having low valence, and positive emotions like happiness and joy having high valence. Arousal represents the level of physiological activation associated with the emotion. Emotions like boredom and calmness have low arousal, whereas excitement and fear have high arousal. The valence-arousal axes create a two-dimensional space divided into four quadrants, with opposing emotions positioned on opposite sides (e.g., joy and sadness).

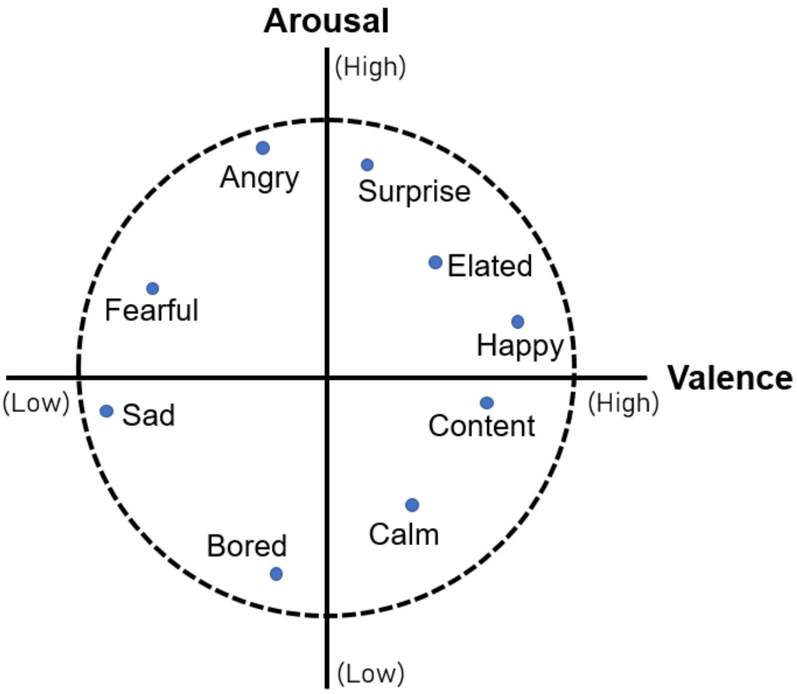


Figure 2.3: An illustration of two-dimensional Russell's circumplex model

The three-dimensional PAD (Pleasure-Arousal-Dominance) model builds on Russell's model by adding a third dimension (see Figure 2.4): dominance [Mehrabian, 1980; Wang et al., 2022; Bota et al., 2019]. Dominance captures the feeling of being in control of an emotional situation. Emotions like pride and anger have high dominance, while shame and fear have low dominance. This additional dimension creates eight sub-spaces or octants, where various emotions are placed depending on their pleasure, arousal, and dominance values. Although the PAD emotion model provides a more nuanced representation of emotional states, Russell's circumplex model remains a more popular emotion model in affective computing research [Wang et al., 2022].

2.1.2 Pain

Williams and Craig [2016] define pain as "a distressing experience associated with actual or potential tissue damage with sensory, emotional, cognitive, and social components". Expressing pain can trigger social reactions such as empathy and care [Williams, 2002]. Understanding the various aspects of pain perception is important for developing machine learning models to detect pain.

Gate Control Theory of Pain

The gate control theory of pain, proposed by Melzack and Wall [1965], provides a framework for understanding how individuals perceive pain. While rooted in neurophysiology, this theory also explains the cognitive and psychological aspects of pain. It suggests that pain signals from the body to the brain are not transmitted directly. Instead, a metaphorical "gate" mechanism in the spinal cord modulates these signals. This gating mechanism

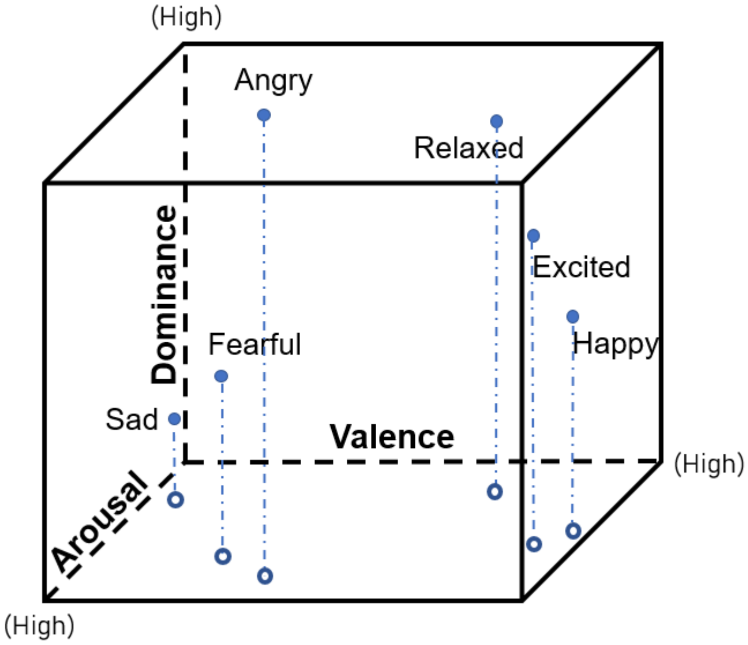


Figure 2.4: An illustration of the three-dimensional PAD model. The solid circles depict the various emotions in the PAD space, and the non-solid circles represent the projection on the Valence-Arousal plane.

can be influenced by various factors, including emotions, cognitive state, and past experiences [Campbell et al., 2020]. This theory explains phenomena like pain being increased by negative emotions [Wiech and Tracey, 2009] and catastrophic thinking hindering recovery [Hadjistavropoulos et al., 2011]. For instance, negative emotions like anxiety can heighten pain perception by further opening the gate, whereas positive emotions (e.g., relaxation) or distraction techniques can reduce perceived pain by closing the gate [Chesterfield PCT Service, 2014].

Diathesis-Stress Model of Pain

The diathesis-stress model has been used in many contexts to explain the effects of stress on the manifestation of illnesses. Turk and Flor [1984] adapted the model for pain, offering a perspective on the link between stress and pain experience. They propose that chronic pain may manifest due to stress, provided three conditions are met: (a) a pre-existing pain-eliciting condition (diathesis) like injuries, (b) recurrent stressful situations (e.g., employment-related stress), and (c) inadequate coping mechanisms. Figure 2.5 illustrates the diathesis-stress model for pain. This model has limitations but it holds particular significance for industrial HRC scenarios. It suggests that the occurrence of pain in workers might not solely be caused by work-related injuries. Understanding this connection could inform the design of cobots that can identify stress and potentially recommend coping mechanisms to human workers, which in turn, might help prevent the development of chronic pain.

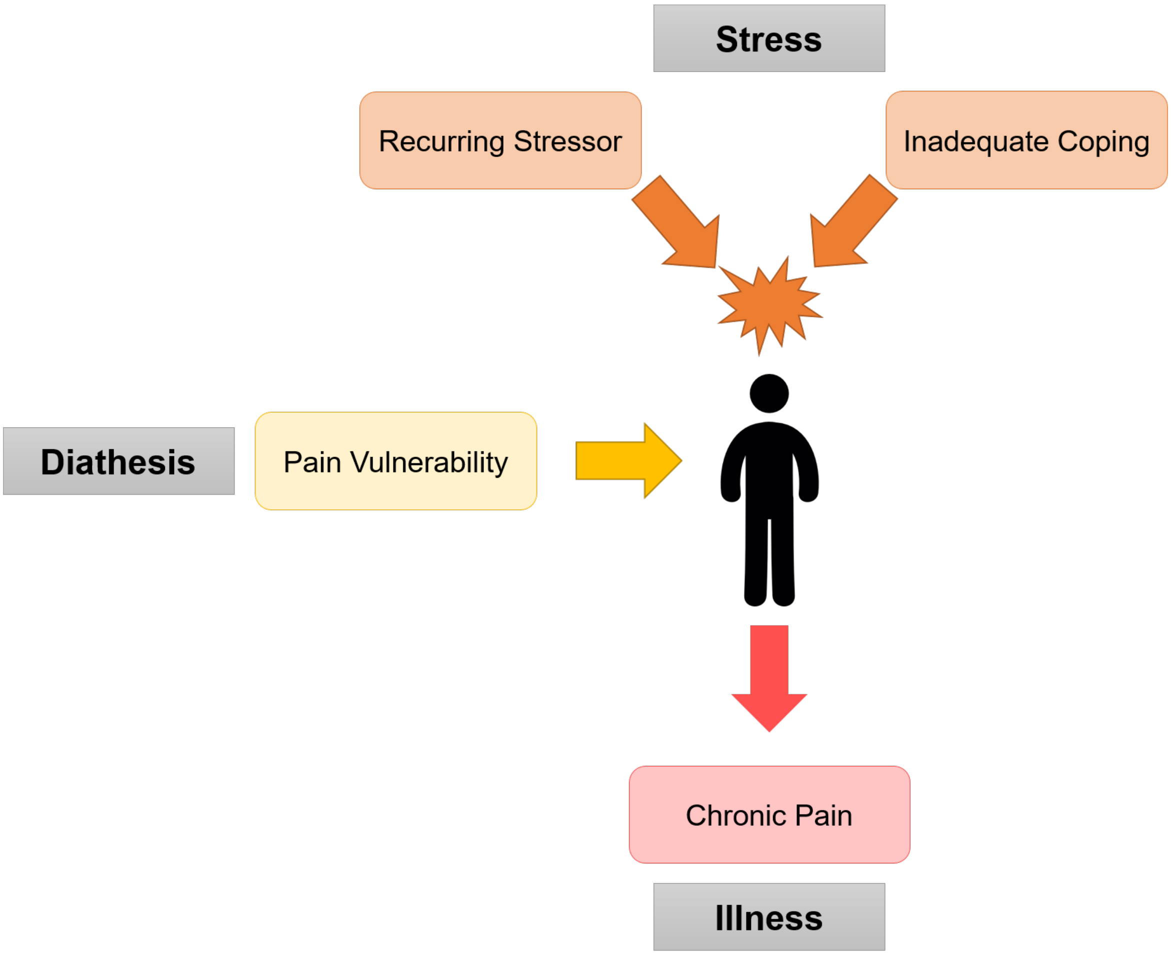


Figure 2.5: Visualization of the diathesis-stress model for pain.

Classifications of Pain

Pain is traditionally classified as acute and chronic, based primarily on the pain duration [Cole, 2002; Lumley et al., 2011]. Acute pain is a short-duration pain, typically lasting less than 30 days. It serves as a warning signal, prompting individuals to take action to prevent tissue damage. On the contrary, chronic pain refers to persistent pain that lasts for more than three to six months. Chronic pain is more complex than acute pain and may cause neurobiological, psychological, and behavioral changes. Sub-acute pain falls between these categories, lasting longer than 30 days but less than three months.

Another classification system, which is gaining traction with the rise of automatic pain recognition models and the availability of public pain datasets, distinguishes between clinical and experimental pain. This classification is based on the type of stimuli that triggers the pain experience [Edens and Gil, 1995; Bouhassira et al., 2003; Kunz and Lautenbacher, 2014]. Clinical pain arises from specific actions (e.g., movement, light touch) in individuals with medical conditions like injuries, surgeries, or medical procedures [Charron et al., 2006]. Experimental pain, on the other hand, is induced under controlled laboratory settings in healthy individuals using stimuli like heat, pressure, or electricity [Edens and Gil, 1995; von Baeyer et al., 2005; Charron et al., 2006]. Traditionally, experimental pain is utilized to study pain experience in a controlled environment [Edens and Gil, 1995; Kim et al., 2004]. Clinical pain can be either acute or chronic depending on the persistence of the stimuli, whereas experimental pain is typically acute.

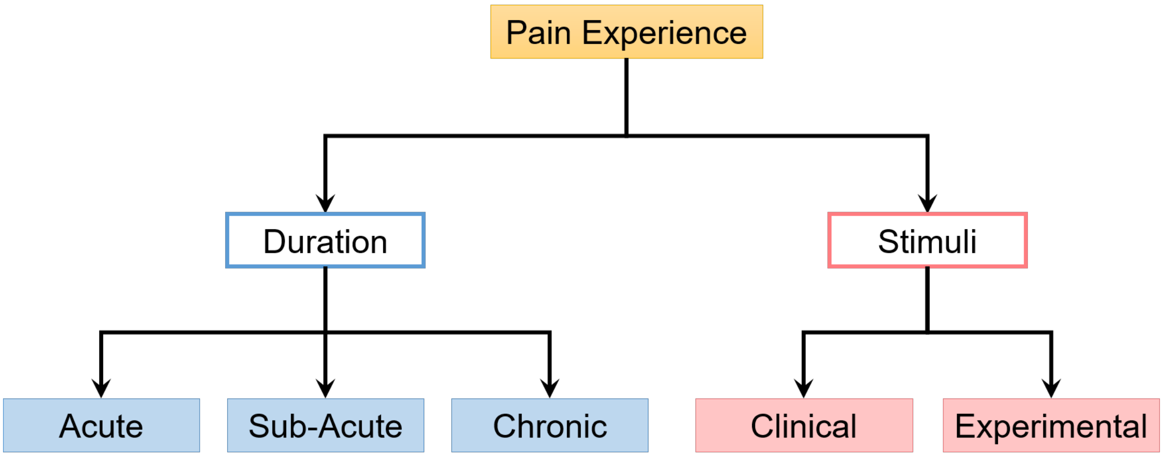


Figure 2.6: Classification of pain based on duration and nature of pain stimuli.

Pain and Emotional States

Defining pain remains a challenge due to its subjective nature. However, a widely accepted definition by the International Association for the Study of Pain (IASP) describes it as “an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage" [Raja et al., 2020]. While there are ongoing discussions for refining this definition, most researchers agree on the presence of an emotional component in pain [Cohen et al., 2018; Rhudy and Meagher, 2001]. Similar to emotions, pain manifests through various behavioral (e.g., facial expressions, body pose) and physiological changes (e.g., elevated heart rate, sweating).

Like many emotions, pain also has distinct facial expressions, with some overlaps. For instance, both pain and disgust (aversive experiences) can involve nose wrinkling [Kunz et al., 2013]. However, considering the entire expression pattern allows for differentiation between pain and other emotions [Simon et al., 2008; Kunz and Lautenbacher, 2014]. Research suggests that pain, focusing on its emotional component, can be mapped onto Russell’s circumplex model as a low-valence (unpleasant) and high-arousal state [Rhudy and Meagher, 2001; Price, 2002; Kyle and McNeil, 2014; Ciuffini et al., 2023]. It is important to note that this categorization considers the emotional aspect and does not imply pain can be solely understood through emotion models.

2.1.3 Stress

Stress is a ubiquitous experience in modern life. While the occurrence of stress is natural, excessive stress can have a significant negative impact on both physical and mental well-being. Although colloquially “stress" is often used interchangeably with other related concepts, the literature distinguishes between stress, stressors, and distress [Wheaton and Montazer, 2010]. Stressor refers to any event, situation, or stimulus that causes strain on an individual. The stress response is the individual's response to the stressor, which may include physiological (e.g., elevated heart rate), psychological (e.g., anger, depression), or behavioral (e.g., smoking) changes [Schwarzer and Schulz, 2003; Khalil and Elfaki, 2014]. Individuals employ coping strategies to mitigate stress. However, when the coping strategies are not effective, stress turns into distress.

Researchers have explored stress from various angles, leading to different theoretical perspectives. Some theories focus on the stressor itself, others on the individual's response, and still others explore the interplay between stimuli, response, and mediating factors like coping strategies [Bailey and Clarke, 1989; Schwarzer and Taubert, 2002; Schwarzer and Schulz, 2003; Shahsavarani et al., 2015; Khalil and Elfaki, 2014]. This categorization broadly classifies stress theories as stimulus-based, response-based, and transactional, respectively. The following paragraphs delve into representative stress theories from each of these categories.

Life-Change Events Model

This model represents a stimulus-based approach to study stress [Khalil and Elfaki, 2014; Wheaton and Montazer, 2010; Schwarzer and Taubert, 2002; Schwarzer and Schulz, 2003; Shahsavarani et al., 2015]. A well-known version of this model was proposed by Holmes and Rahe [1967], who developed the Social Readjustment Rating Scale (SRRS) as a tool to measure major stressors over a certain period (e.g., over the last year). This model posits that stress arises from the accumulation of stressful life events (stressors). The SRRS considers 43 stressful life events, each measured in life-change units. Higher values are assigned to events like the death of a spouse/loved one, and lower values for events like minor violations of law. The underlying assumption for the different scores is that some events require more effort to overcome than others [Schwarzer and Taubert, 2002; Shahsavarani et al., 2015]. Moreover, some positive events are also included in the SRRS list as any change can be stressful, regardless of whether it is positive or negative [Schwarzer and Schulz, 2003]. Individuals with higher SRRS scores (indicating they experienced major life changes) are considered more likely to develop stress-related health problems.

Researchers have identified multiple limitations of this model. Critics point out that it does not account for variances in individual appraisal of various situations [Schwarzer and Taubert, 2002; Shahsavarani et al., 2015; Khalil and Elfaki, 2014]. For instance, individuals may perceive and respond differently to the stressor of minor violations of law. Additionally, the model can assign similar scores to individuals who experience very different life events. For example, someone who experienced the death of a spouse might have a similar SRRS score to someone who got married and was fired from work. However, questions have been raised on whether these experiences can be considered equal and analyzed in a similar way [Schwarzer and Schulz, 2003].

General Adaptation Syndrome (GAS)

The GAS model, proposed by Selye [1950], exemplifies a response-based approach to studying stress [Khalil and Elfaki, 2014; Schwarzer and Taubert, 2002; Schwarzer and Schulz, 2003; Shahsavarani et al., 2015]. This model focuses on the physiological stages the body

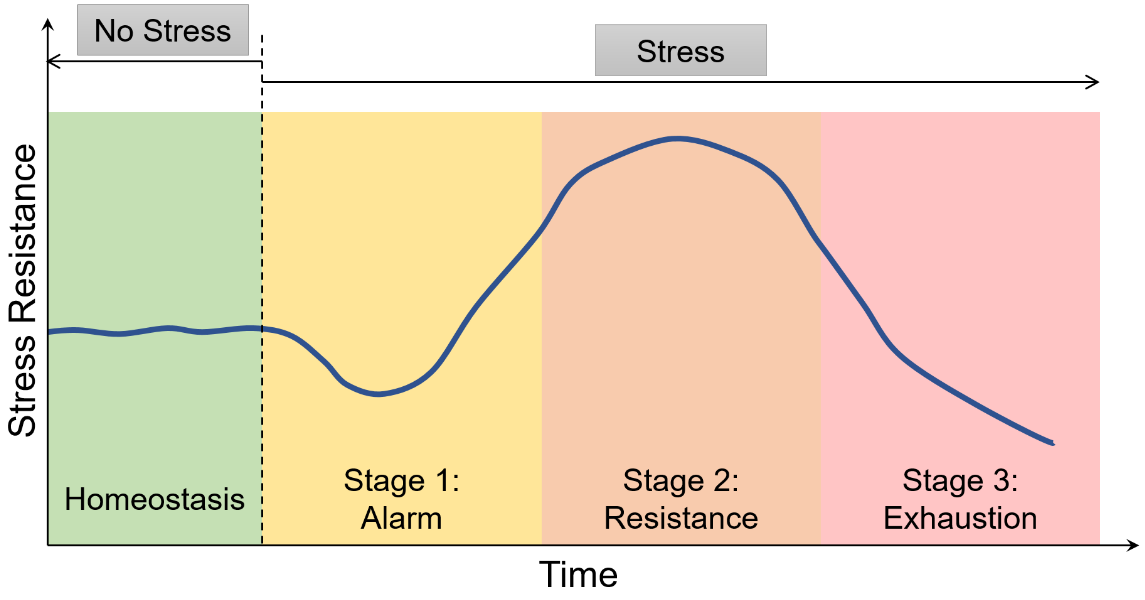


Figure 2.7: Illustrative plot of the stress resistance over time in the GAS stress model.

undergoes in response to a stressor. According to Selye, stress experience develops three stages (see Figure 2.7): alarm reaction, resistance, and exhaustion. The alarm reaction refers to the body's initial response, also called the "fight-or-flight" response. This stage triggers physiological changes such as elevated heart rate, blood pressure, and breathing rate. If the stressor persists, the body enters this second stage, i.e., the resistance stage, where it adapts to the demands of the situation. This stage utilizes bodily resources for adaptation and repairing damages caused during the alarm reaction stage. If the stressor remains unaddressed and resources gets depleted, it enters the exhaustion stage. This stage can lead to stress-related illnesses, burnout, depression, etc. The GAS model highlights that stress leads to "defense [adapting response] and damage".

Based on this model, Selye [1976] formulated a generic definition of stress as the body's "non-specific response to any demand". While this model holds value in understanding physiological responses to stress, it has limitations. Psychology researchers point out that GAS does not consider the cognitive and emotional aspects of the stress experience Schwarzer and Taubert [2002]; Schwarzer and Schulz [2003]. Additionally, similar to the SRRS model, GAS does not account for individual differences in stress perception.

Transactional Model of Stress and Coping

This model was proposed by Lazarus [1966] and refined in a subsequent work [Lazarus and Folkman, 1987]. This model represents a transactional approach, viewing stress as a dynamic interplay between the individual and the environment [Khalil and Elfaki, 2014; Schwarzer and Schulz, 2003; Schwarzer and Taubert, 2002; Shahsavarani et al., 2015; Dewe et al., 2012]. It addresses the limitations of previous models like the GAS model and SRRS by emphasizing the role of cognition and appraisal in shaping the stress experience.

There are three core elements of this model: transaction, process, and emotional system. Transaction or relationship refers to the specific encounter between the individual

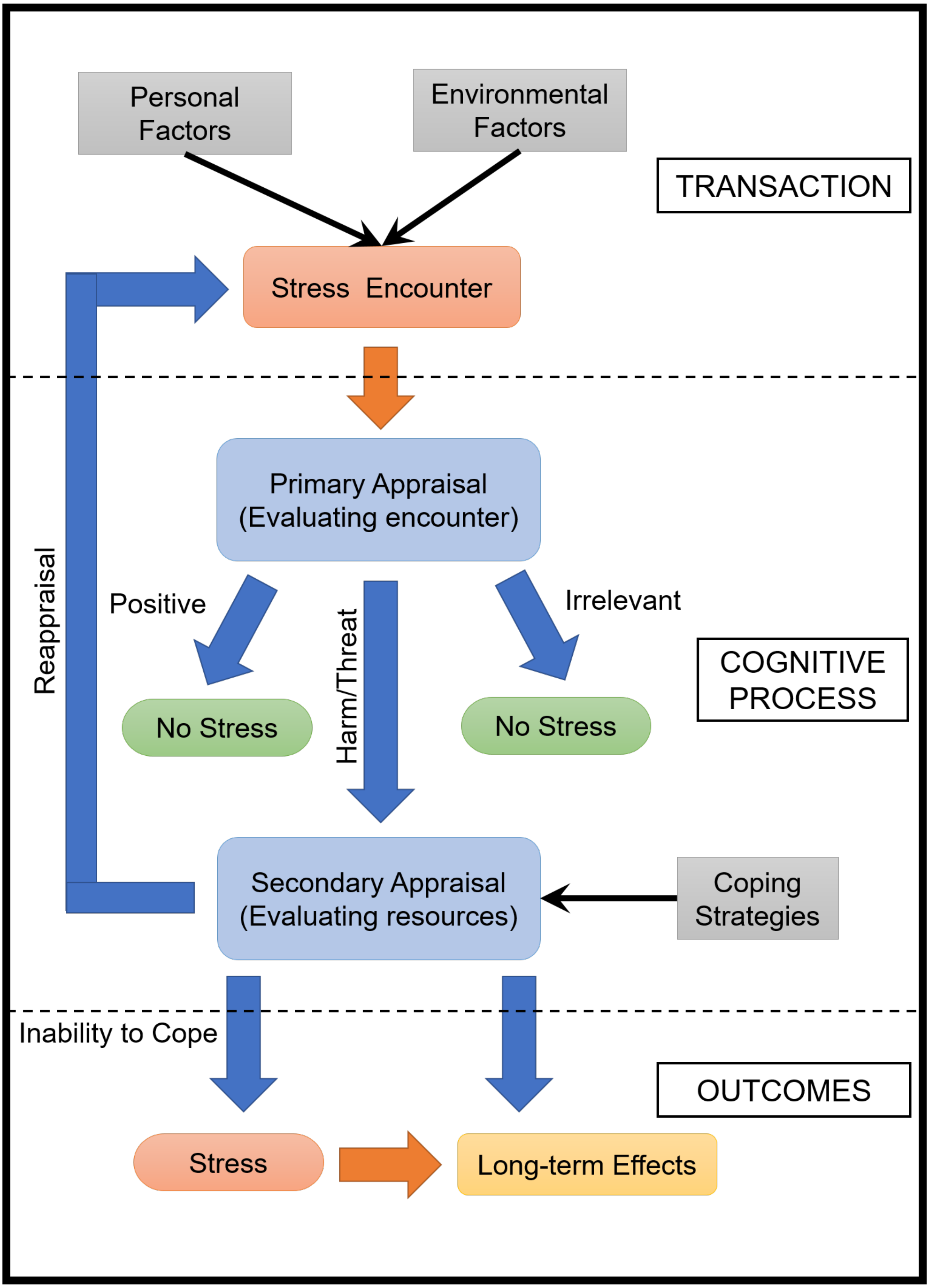


Figure 2.8: A depiction of the transactional model of stress and coping.

and the environment. A threat (stressor) gains meaning only in context with the individual experiencing it. In other words, stress is not inherent to the situation itself but arises from the interaction between the person and the environment. The process element emphasizes the dynamic nature of stress. The stress experience is constantly evolving as the individual and their environment adapt. The transaction and process elements supplement an emotional system view of the stress experience. The emotional system consists of causal antecedents, mediating processes, and outcomes. Causal antecedents are person-related factors (e.g., beliefs, skills, sense of control) and environment-related factors (e.g., demands, resources, constraints). Mediating processes refer to an individual's cognitive appraisal of the situation and coping strategies. The outcomes are the consequences of the stress experience, including both short-term (e.g., physiological changes, affective experience) and long-term effects (e.g., psychological well-being, illnesses).

A crucial aspect of this model is the role of mediating processes, particularly cognitive appraisal and coping processes. Lazarus and Folkman identified two types of cognitive appraisals: primary and secondary. Primary appraisal, also called demand appraisal, refers to the initial evaluation of the situation to determine if it is irrelevant, positive, or stressful (threat, harm, or challenge) [Schwarzer and Schulz, 2003; Khalil and Elfaki, 2014; Dewe et al., 2012]. During the secondary appraisal or resource appraisal, the individual evaluates their coping resources and options for dealing with the stressor. Coping efforts are continuously adjusted based on re-evaluations of the situation. The model also proposes two coping strategies: problem-focused coping (actively managing the situation) and emotion-focused coping (regulating emotions arising from the situation). However, later studies have proposed additional strategies such as meaning-centered and relationship-social coping strategies [Dewe et al., 2012].

A visualization of the various aspects of this transactional model of stress and coping is presented in Figure 2.8. This model has led to a widely accepted definition of stress: "a particular relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being" [Lazarus and Folkman, 1984]. This definition highlights the subjective nature of the stress experience.

Classifications of Stress

Similar to pain, stress can be classified into three categories based on the duration of the stressor: acute, chronic intermittent, and chronic stress [Kovács et al., 2005; Shahsavarani et al., 2015; Giannakakis et al., 2019]. Acute stress refers to short-term stress, typically involving a single encounter with the stressor. Examples include a public speaking event or cold pressor stress. Chronic stress refers to long-term stress, where the stressor is continuously present (e.g., chronic work overload, poor financial situations). Chronic intermittent stress or repeated stress refers to a situation where an individual experiences a stressor repeatedly for a prolonged duration. For example, a student facing multiple exams throughout an exam week has repeated encounters with the stressor (exam). While all types of stress can elicit physiological changes, chronic stress can lead to stress-related illnesses such as cardiovascular diseases, depression, and sleep issues.

Stress can also be categorized based on the nature of the stressor into physiological and psychological stress [Lu et al., 2021; Shahsavarani et al., 2015; Giannakakis et al., 2019]. Physiological stress arises from physical stimuli that disrupt the body's balance (homeostasis). This class of stressors can include heat stress, loud noises, or pain from an injury or illness. Psychological stress, on the other hand, is induced by an individual's thoughts, emotions, and perceptions of the situation. It can be further divided into four subcategories: emotional stress (anxiety, fear, etc.), cognitive stress (information overload, interruptions, etc.), perceptual stress (competition, perceptual workload, etc.), and psychosocial stress (e.g., social evaluation, social defeat, social confrontation). Unlike physiological stress, which may have a localized source (e.g., pain in a specific body part), psychological stress is more intangible.

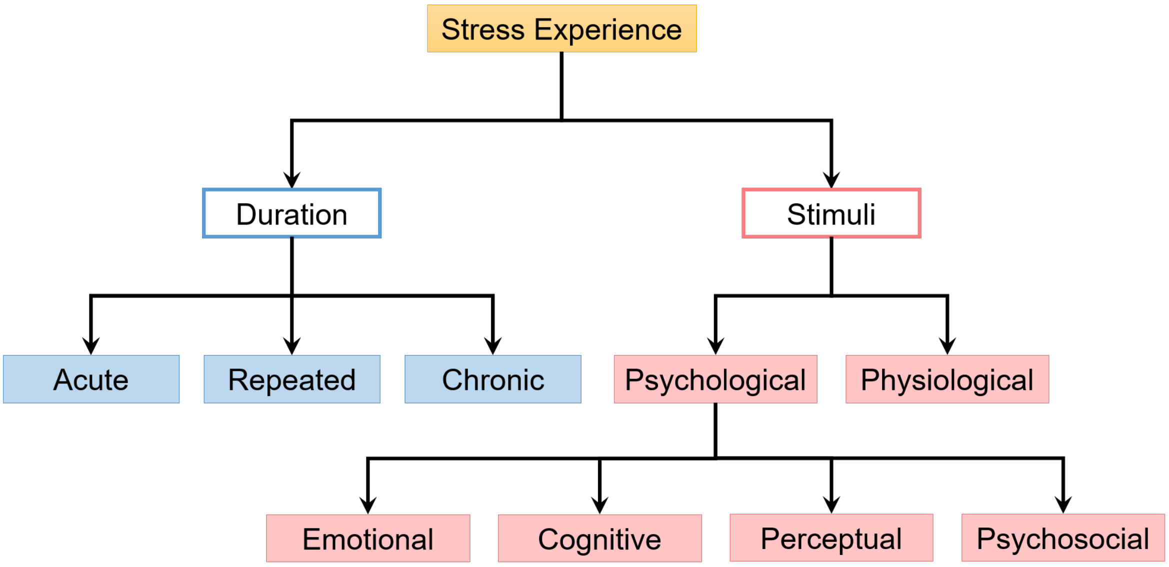


Figure 2.9: Classification of stress based on duration and nature of the stressor.

2.1.4 Flow

The concept of flow state was first described by Csikszentmihalyi [1975] and refers to a state of deep engagement and optimal experience. This state is often associated with enhanced performance, positive emotions, and improved mental well-being [Norsworthy et al., 2021; Jackson et al., 2001; Peifer et al., 2022]. Understanding the characteristics and manifestations of the flow experience is important for developing machine learning models that can facilitate the occurrence of this state.

Characteristics of Flow

There are nine characteristics of the flow state, which are widely accepted in the literature [Jackson and Csikszentmihalyi, 1999; Nakamura and Csikszentmihalyi, 2002; Jackson et al., 2001; Csikszentmihalyi, 2020; Norsworthy et al., 2021] and are briefly described below (adapted for the context of industrial HRC).

- **Challenge-Skill balance**: The perceived challenge level of the task matches the worker's perceived skill level.
- **Clear goals**: The purpose of the task and the steps to accomplish it are clearly defined and understood by the worker.
- **Unambiguous feedback**: The worker receives clear and immediate feedback on their task performance.
- **Merging of action and awareness**: The worker becomes deeply involved in the task, with their actions becoming automatic.
- **Concentration on the task at hand**: The worker's attention is completely focused on the present moment and the task at hand, with no distracting thoughts or mind-wandering.
- **Sense of control**: The worker feels a sense of control or agency over their actions, the task, and the situation.
- **Loss of self-consciousness**: The worker becomes less worried about themself and less concerned with how they are perceived by others.
- **Transformation of time**: The worker's subjective experience of time is altered.
- **Autotelic experience**: The word autotelic is derived from the Greek words auto meaning self and telos meaning purpose. This characteristic implies that the worker finds the task intrinsically rewarding and enjoyable for its own sake, rather than for external rewards (e.g., monetary bonuses).

Out of the nine characteristics, the first three (challenge-skill balance, clear goals, and unambiguous feedback) are considered the pre-conditions for the flow state to occur, whereas the other six are considered descriptions of the state itself [Norsworthy et al., 2021; Peifer et al., 2022]. The challenge-skill balance is considered a primary characteristic as research has found correlations between this characteristic and the other eight. This has led to the utilization of challenge-skill balance as a means to elicit a flow state during tasks.

Three-channel Flow Model

The most simple representation of the flow experience is Csikszentmihalyi's three-channel flow model, which captures boredom, anxiety, and flow states [Csikszentmihalyi, 1975; Pearce, 2005; Peifer et al., 2014]. In psychological flow models, the relationship between skills and challenges is utilized to map the corresponding experience states or “channels". As illustrated in Figure 2.10, this model depicts channels as zones on a graph, visualized with respect to a diagonal line (x = y). This diagonal line represents a state of equilibrium where perceived challenge and skill levels are equal. The flow state is achieved when the perceived challenge of a task matches the individual's perceived skill level. This state is represented by the zone closest to the central diagonal line (identity line).

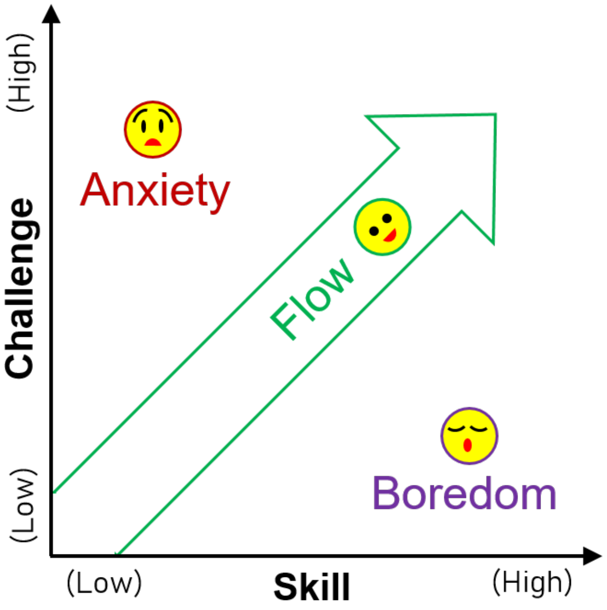


Figure 2.10: An illustration of Csikszentmihalyi's three-channel model, showing how perceived challenge and individual skill influence the experience of boredom, anxiety, and flow.

Deviations from this ideal balance can lead to negative experiences. When the challenge of the task exceeds the individual's perceived skill level (zone above the identity line), feelings of anxiety and stress arise. Conversely, when the challenge is significantly lower than the skill level (zone below the identity line), boredom and disengagement become prevalent.

Four-channel and Eight-channel Flow Models

The three-channel flow model provides a foundational framework, but some researchers have proposed models with additional states to capture a wider range of experiences. The four-channel model [Csikszentmihalyi, 1975] extends the original model to include the state of apathy (see Figure 2.11). In this model, the four states (flow, boredom, apathy, and anxiety) are represented as four quadrants formed by the challenge-skill axes [Jonsson and Persson, 2006; Lambert et al., 2013]. The newly added state of apathy is characterized by low challenge and low skill levels. Boredom (low challenge, high skill) and anxiety (high challenge, low skill) are on the opposite quadrants. Flow is characterized by high challenge and high skill in this model.

Both three-channel and four-channel flow models distinguish only flow as a positive state, with all others considered negative [Jonsson and Persson, 2006]. The eight-channel model or experience fluctuation model, proposed by Massimini et al. [1987], presents a more nuanced representation of various experiences [Lambert et al., 2013]. As depicted in Figure 2.12, the model maps these states within a circular formation divided into eight 45-degree sectors on a two-dimensional challenge-skill plane. Unlike previous models, the eight-channel model considers not just high and low, but also moderate levels of perceived challenge and skill to differentiate the various states.

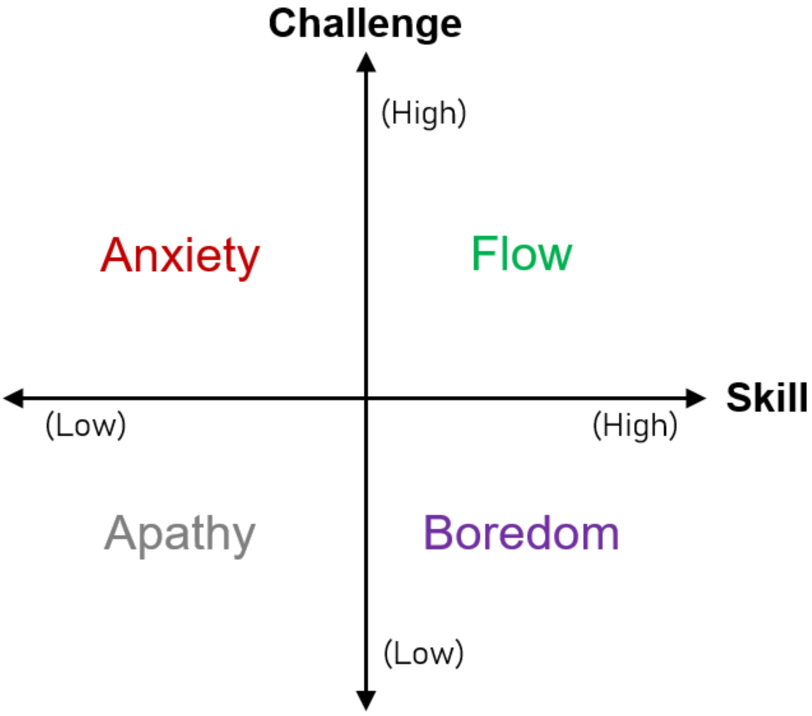


Figure 2.11: An illustration of the four-channel flow model, showing various experiences in terms of perceived challenge and individual skill levels.

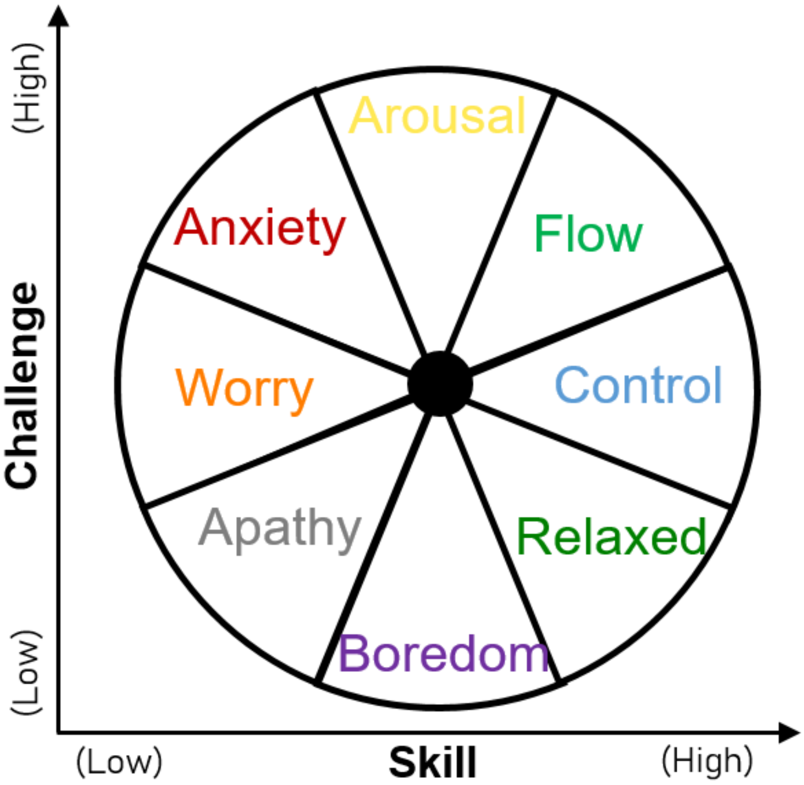


Figure 2.12: An illustration of the eight-channel flow model, showing various experiences in terms of perceived challenge and individual skill levels.

2.2 Behavioral Signals

This section presents a background for some of the facial modalities and cues that are leveraged in this thesis. It is important to note that some modalities are better suited than others in detecting certain states. So, this section also briefly discusses these signals in the context of their potential for detecting specific worker states.

2.2.1 Facial Images/Videos

Facial images and videos represent a rich source of behavioral patterns such as emotional expressions and head orientation. These data are typically recorded using an RGB camera facing a person. Machine learning models can leverage these cues to detect worker states that may be visible through their faces (e.g., pain).

Action Units

Facial expressions are one of the most informative modalities of non-verbal communication [Ko, 2018; Valstar et al., 2017]. Facial expressions are produced by contractions and relaxation of facial muscles, which result in momentary changes in facial appearance. These movements are known as Action Units (AUs). Ekman and Friesen [1978] devised the Facial Action Coding System (FACS) based on AUs, assigning them specific numbers. The system has numbered AUs for capture various movements, including facial muscles (e.g., AU 6 - cheek raiser, AU 12 - lip corner puller), eye movements (e.g., AU 62 - eyes turn right, AU 64 - eyes down), and head orientations (e.g., AU 55 - head tilt left, AU 53 - head up) [iMotions, 2022]. The intensity of each AU is scored on a 5-point scale from A (trace) to E (maximum) based on how pronounced the movement is [Clark et al., 2020]. Specific AUs or combinations of AUs are associated with different facial expressions and corresponding emotions. For example, the prototypical expression of happiness involves simultaneous activation of AU 6 and AU 12. Emotional expressions typically use 30 facial muscle AUs, 12 from upper face and 18 from lower face muscles [Tian et al., 2001]. A comprehensive list of AUs and their occurrences in Ekman's basic emotions are visualized in iMotions website [iMotions, 2022].

Some limitations of using AUs as features have been identified in the literature [Clark et al., 2020]. FACS captures visible changes in facial movement and doesn't account for subtle visible changes (e.g., changes in muscle tone). Moreover, it does not capture other facial changes like sweating (e.g., during stress) or slight changes in skin complexion (e.g., during anger).

Facial Landmarks

Facial landmarks are distinct and identifiable locations on the face, such as the corners of the eyes, the tip of the nose, and the corners of the mouth. Each location or key point is represented using two values, indicating the x and y coordinates. The number of key points depends on the available annotations in a dataset, i.e., unlike FACS for AUs, there is no standard landmark system. The minimum number of key points in public facial landmark datasets can be as low as four [Wu and Ji, 2019; Johnston and de Chazal, 2018]. However, the number of annotated key points has increased in recent datasets [Wu and Ji, 2019], with some datasets often using 68 key points. Facial landmarks serve as reference points for localizing and tracking facial features, enabling tasks like face cropping, facial alignment, and head pose estimation.

Emotion Expression

Inferring emotions from facial expressions has been extensively studied in affective computing [Wang et al., 2022; Ko, 2018; Canal et al., 2022]. Some works utilize extracted features (e.g., AUs, facial landmarks) to recognize the expressed emotion. Others employ deep learning methods to predict the emotions directly from the facial images/videos.

While Ekman's principles of basic emotions argue for universal facial expressions based on studies across cultures, others point to significant variability in how emotions are facially expressed and perceived [Klingner and Guntinas-Lichius, 2023; Stahelski et al., 2021]. For example, a scowling expression could be interpreted as anger by some people and disgust by others. This has led to mislabelling in many datasets. Notably, the mislabelling occurs predominantly among negative emotions, particularly between anger and disgust [Stahelski et al., 2021]. This has led to a shift towards estimating the valence-arousal of an expression rather than a classification approach.

Pain Expression

Unlike emotions, research has found similarities in the perception of pain expressions between different cultures (e.g., Western vs. Eastern cultures) [Klingner and Guntinas-Lichius, 2023]. Moreover, facial expressions are considered reliable for detecting pain as they involve certain muscles, especially around the eyes, that cannot be voluntarily controlled [Williams, 2011; Hadjistavropoulos et al., 2011]. In other words, pain expressions are difficult to completely mask or fake.

The development of FACS invigorated the goal of establishing a standard expression of pain [Prkachin, 2009]. Many researchers studied the various AUs activated during pain. For example, Prkachin [1992] studied facial expressions during pain induced through multiple stimuli (cold, pressure, ischemia, and electricity) and found four patterns consistent across all four stimuli: brow lowering, nose wrinkling, lid tightening, and eye closure. Although AUs like lip-corner pulling and blinking were observed in some instances, they were not as consistent as the four "core" patterns. Similarly, Kunz et al. [2019] investigated existing pain datasets and found eye closure was more associated with clinical pain than experimental pain. However, they found mouth opening along with the other three core patterns were consistent across both types of pain. These findings suggest that developing a generic pain detection model is a feasible goal.

2.2.2 Gaze

Like facial expressions, gaze is a visual signal people use in their communications. However, the role of gaze changes depending on the context [Hamilton, 2016; Frischen et al., 2007]. For example, although a direct gaze can indicate attentiveness, an averted gaze does not always indicate distraction as it can be part of expressing emotions like embarrassment. Moreover, gaze can be perceived as positive (e.g., love/attraction), neutral (e.g., regulating turn-taking), or negative (e.g., staring). Gaze is especially interesting in HRC scenarios as it can enhance collaboration by facilitating aspects such as resolving ambiguities and establishing joint attention [Mehlmann et al., 2014; Schneider and Pea, 2017; Mitev et al., 2018].

Gaze is typically recorded using standard RGB cameras, remote eye-tracking devices, or wearable gaze trackers. Most dedicated eye trackers (wearable and remote) use infrared (IR) illuminators along with IR cameras to track the eyes and pupil movements [Caporusso et al., 2022; Karmakar et al., 2024]. The IR light creates bright pupil and corneal reflections that are easier to detect and track than visible light. Moreover, their high framerate and resolution lead to more accurate gaze-tracking than standard cameras. However, RGB cameras are a cost-effective and ubiquitous alternative. In addition, standard cameras are non-obtrusive compared to wearable trackers.

There are primarily three aspects of gaze that are used to derive relevant features: saccade, fixation, and gaze direction [Mézière et al., 2021; D'Angelo and Schneider, 2021]. Saccades refer to rapid eye movements that change the gaze from one location in the field of view to another. Fixations are the periods between saccades when the eyes remain relatively still and focused on a specific location. Gaze direction is the location at which the person is looking.

Gaze Direction

The first step before inferring saccades and fixations is determining the gaze direction. This task is also called gaze estimation. Gaze is estimated either as a two-dimensional or a three-dimensional vector. The two-dimensional representation uses the pupil's angular coordinates (pitch, yaw), whereas the three-dimensional vector also includes the depth information [Kwon et al., 2006; Yu and Odobez, 2020].

When using standard camera images, gaze direction can be inferred using two types of information: eye gaze and head pose [Matsumoto et al., 2000; Cheng et al., 2024]. The eye-based gaze estimation is more accurate than the estimation based on head pose [Palinko et al., 2016]. However, the head pose is more robust when occlusions are (partially) covering the eyes.

Attention and Distraction

Attention and distraction detection have become increasingly important in many domains, including driver safety, education, and human-machine interaction [Cartella et al., 2024]. In this thesis, attention refers to visual attention, i.e., the cognitive process of selectively focusing on specific areas of the visual environment while ignoring other perceivable information. Distraction occurs when the visual attention is on areas unrelated to the task or HRC scenario (e.g., interruptions) [Kotseruba and Tsotsos, 2022]. In such use cases, gaze direction is often considered a measure of attention to specific areas of interest in the field of view. Subsequently, machine learning models are developed to map the individual's gaze direction to areas or objects within the visual field.

Social Gaze Cueing

There are many social contexts in which people employ gaze-based cues [Hamilton, 2016; Frischen et al., 2007]. Understanding all the social roles of gaze cues is a broad research area. However, some of the social gaze cues are interesting for HRC research. One such cue that has gained significant traction is recognizing intent [Belardinelli, 2023]. Studies have explored inferring the operator's intention (e.g., the next assembly step, component choices) from their gaze to improve efficiency during collaboration. The core idea behind gaze-based intention recognition is that people tend to look at objects or areas they plan to work on before starting work on them.

Another interesting phenomenon is attention orienting through gaze cues. Research has shown that humans can use gaze cues to direct others' attention in a shared space [McKay et al., 2021; Edwards et al., 2015]. This is often seen as a necessary step in initiating joint attention. Moreover, gazing at someone can capture their attention [Frischen et al., 2007; Akechi et al., 2013; Lee et al., 2020]. All the discussed scenarios highlight the need for the cobot to monitor an individual's gaze for social cues during collaborative tasks.

2.3 Physiological Signals

This section describes the physiological signals utilized in this thesis and how they reflect some of the relevant worker states discussed in this thesis.

2.3.1 Autonomic Nervous System

The human nervous system consists of two main parts: central and peripheral [Thau et al., 2022; Guy-Evans, 2023]. The central nervous system consists of the brain and spinal cord, whereas the peripheral nervous system consists of nerves and ganglia. Further, the peripheral nervous system can be divided into the somatic (responsible for voluntary skeletal muscles) and the autonomic nervous system (ANS). The ANS regulates involuntary bodily functions and maintains homeostasis in the body [McCorry, 2007; Guy-Evans, 2023; Waxenbaum et al., 2023; Bota et al., 2019]. It controls vital functions like heart rate, breathing, and digestion.

The ANS has two components: sympathetic, parasympathetic, and enteric nervous system. Activation of the sympathetic system leads to a “fight-or-flight" response to a perceived strenuous situation. This response is characterized by elevated activity and attentiveness, resulting in physiological changes such as increased heart rate, sweating, and respiration. It also slows lower-priority processes like digestion. On the contrary, the parasympathetic nervous system activation facilitates a “rest-and-digest" response during relaxation periods. This response is associated with decreased bodily activity (e.g., lower heart rate and blood pressure). The enteric system is independent of the other parts of the

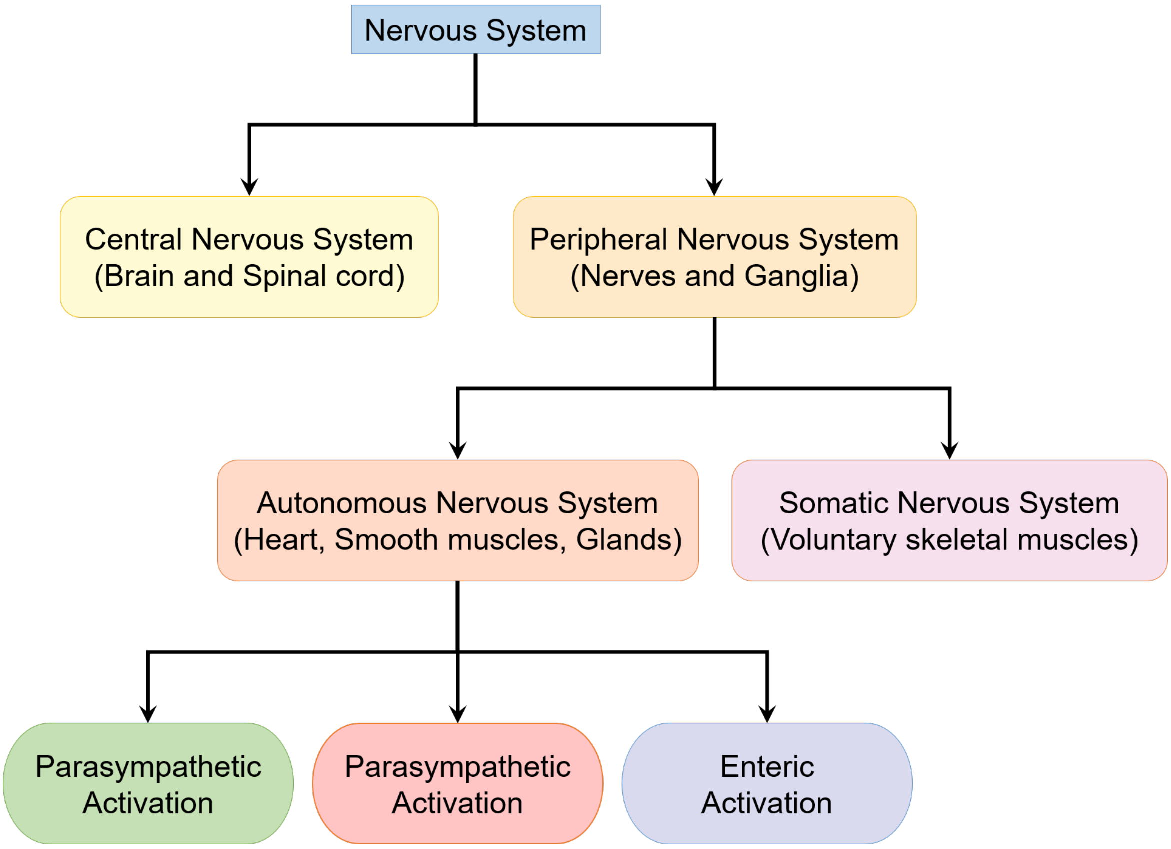

Nervous System
Central Nervous System (Brain and Spinal cord)
Peripheral Nervous System (Nerves and Ganglia)
Autonomous Nervous System (Heart, Smooth muscles, Glands)
Somatic Nervous System (Voluntary skeletal muscles)
Parasympathetic Activation
Parasympathetic Activation
Enteric Activation

Figure 2.13: Diagram showing the division of the various components of the nervous system.

nervous system. It mainly facilitates the movement of water and electrolytes across the intestinal wall and coordinates the gut muscles to produce peristalsis.

In summary, the sympathetic modulations of the different organs and cells are prevalent during high arousal states, whereas parasympathetic modulations are predominant during relaxation.

2.3.2 Heart Rate Variability

Heart rate variability (HRV) represents the variation in time intervals between consecutive heartbeats. It reflects the activations of the sympathetic and parasympathetic branches of the autonomic nervous system.

The HRV values can be computed from signals capturing heartbeats, i.e., ECG or BVP. Brief descriptions of these signals are presented below.

ECG

An ECG device records the electrical activity of the heart over time. The ECG signal can be recorded in a non-invasive manner by placing electrodes on the skin of the chest, arms, and legs. These electrodes detect the minor electrical changes when the heart muscles depolarize and repolarize during each heartbeat [Gacek, 2011; Singh and Krishnan, 2023]. The electrical signals are amplified and recorded as a series of repeating waveforms, forming the ECG signal. Typically, ECG signals are recorded at a high frequency of 500 - 1000

Hz [Bota et al., 2019].

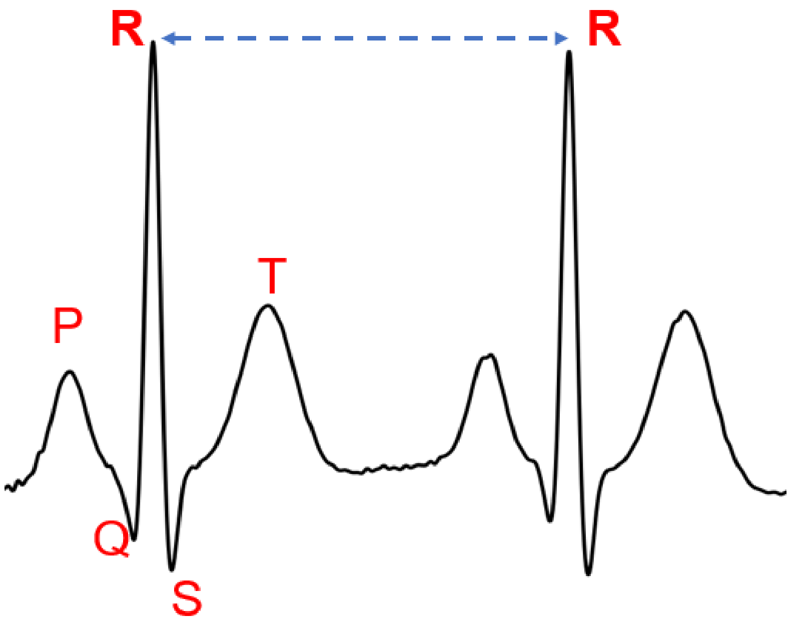


Figure 2.14: An illustrative plot of an ECG waveform with two beats, labeled with the P-QRS-T waves. The blue dotted line represents the R-R distance (an NN interval).

The heart has two atria (right and left) that perform blood collection and two ventricles (right and left) that pump the oxygenated blood to the rest of the body [Singh and Krishnan, 2023; Al-Qazzaz et al., 2014]. These activities are represented in the ECG signal as characteristic waves. An ECG waveform consists of P-QRS-T waves, where the P wave represents atrial depolarization, the QRS complex represents ventricular depolarization, and the T wave represents ventricular repolarization [Gacek, 2011]. While the shape and amplitude of these waves provide valuable information, the calculation of HRV involves identifying the timestamps of the peak of the QRS complex. The time interval between two consecutive peaks is called the R-R distance or NN (normal-to-normal) interval. The NN intervals are computed over a period (e.g., 1 minute, 5 minutes, etc.) to obtain the HRV signal. Figure 2.14 shows a snippet from a sample ECG signal, marked with the P-QRS-T waves and measurement of the NN interval.

BVP

The BVP signal represents the dynamic changes in blood volume in the peripheral tissues, such as the fingertips or face [Peper et al., 2010; Bota et al., 2019]. The blood volume changes are caused by the rhythmic contraction and relaxation of the heart during each cardiac cycle. Photoplethysmography (PPG) is the most common method to record BVP [Peper et al., 2010; Yu et al., 2018]. A PPG sensor shines light onto the skin and detects the changes in light absorption or reflection due to the changes in blood volume. The BVP signals are typically recorded at frequencies lower than 100 Hz [Bota et al., 2019].

The BVP signal waveform typically consists of the systolic peak, diastolic trough, and dicrotic notch. While the systolic peak corresponds to the maximum blood volume during ventricular contraction, the diastolic trough corresponds to the minimum blood volume during ventricular relaxation [Al-Qazzaz et al., 2014]. In ECG, the terms depolarization and repolarization were used to represent contraction and relaxation. Depolarization and repolarization indicate the electrical phenomenon, whereas contraction and relaxation are mechanical terminology [Mammen et al., 2004]. The dicrotic notch is a small deflection or secondary peak that appears on the descending limb of the BVP waveform (after the primary systolic peak). It is caused by the closure of the aortic valve at the end of ventricular systole [Li et al., 2003]. Similar to ECG, time intervals between two consecutive systolic peaks are computed to obtain the HRV signal. The various components of a BVP signal are illustrated in Figure 2.15, along with the visualization of the computation of NN intervals from the BVP signal.

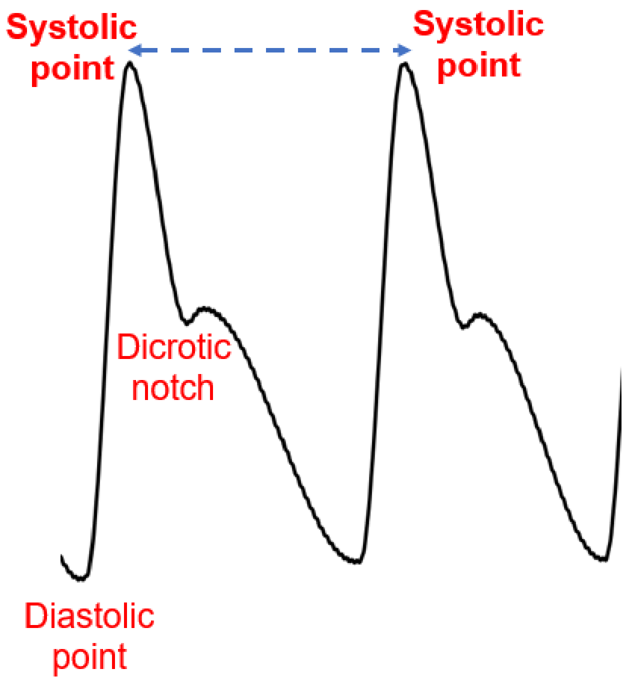


Figure 2.15: An illustrative plot of a BVP waveform with two beats, labeled with the systolic peak, diastolic point, and dicrotic notch. The blue dotted line represents the interval between two systolic peaks called an NN interval.

Although both ECG and BVP measure heart activity and are often highly correlated, the ECG signal is deemed more suitable for computing HRV [Yu et al., 2018]. A plausible reason is the sharp QRS peaks in ECG are more accurately detected than the curvy systolic peaks in BVP. Moreover, the quality of the BVP signal is susceptible to various factors, including skin pigmentation, motion artifacts, and environmental conditions (e.g., lighting) [Bota et al., 2019].

HRV Features

HRV is widely recognized as a biomarker for the ANS modulations [Shaffer and Ginsberg, 2017; Kim et al., 2018b; Arakaki et al., 2023]. Consecutively, handcrafted HRV features and their link to sympathetic and parasympathetic activations have been explored in the literature. HRV features are broadly classified as time domain, frequency domain, and non-linear features. Time domain features are statistical values computed from the NN intervals. The frequency domain features are calculated from the power spectral density analysis of the HRV signal. The non-linear features include values obtained from entropy and poincaré plot analyses. Some of the frequently studied HRV features from each domain and how they reflect the ANS activations are described below.

- **Mean NN**: This is the simplest time domain feature, which represents the average time interval between two beats of a given HRV segment. It can be computed as:

  $$meanNN = \frac{1}{N}\sum_{i=1}^{N} NN_i$$

  Here, NN_i refers to the i^{th} NN interval and N is the total number of NN intervals in the HRV signal.

  The mean NN decreases with sympathetic activation and increases with parasympathetic activation [Kim et al., 2018b; Peabody et al., 2023; Purnamasari et al., 2019].

- **SDNN**: This feature is the standard deviation of NN intervals and is computed as:

  $$SDNN = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(NN_i - meanNN)^2}$$

  SDNN is a measure of the total variability of the HRV signal. This feature reflects the sympathetic activity as a lower SDNN is an indicator of sympathetic modulations [Shaffer and Ginsberg, 2017; Purnamasari et al., 2019; Arakaki et al., 2023]. Similarly, parasympathetic modulations lead to larger HRV and higher SDNN values [Peabody et al., 2023; Shaffer and Ginsberg, 2017].

- **RMSSD**: It is short for root mean square of successive differences (RMSSD). This is another popular time domain feature, especially in stress detection. It captures the variations in adjacent NN intervals and is calculated as:

  $$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1} {SD_i}^2}$$

  SD_i refers to the i^{th} successive difference and $SD_i = NN_{i+1} - NN_i$. For a given HRV segment, there are $N - 1$ successive differences.

  The RMSSD values reflect the parasympathetic activity, i.e., increased parasympathetic modulations lead to larger HRV and higher RMSSD [Kim et al., 2018b; Peabody et al., 2023; Shaffer and Ginsberg, 2017].

- **pNN50**: This is another feature that quantifies variations in adjacent NN intervals. Specifically, it calculates the number of adjacent NN intervals that differ by more than 50 milliseconds and represents them as a percentage of all such successive intervals.

$$pNN50 = \frac{\text{num of } SD_i > 50ms}{N-1} \times 100$$

Like RMSSD, pNN50 also represents parasympathetic influence, with a higher value representing parasympathetic activation [Kim et al., 2018b; Shaffer and Ginsberg, 2017].

- **LF**: This feature represents the power intensity in the low-frequency band (0.04 Hz to 0.15 Hz) of the HRV signal. The frequency features of HRV are computed from power spectral density, which is estimated by applying techniques like fast Fourier transform or autoregressive modeling on the HRV signal [Purnamasari et al., 2019]. The LF feature is represented as:

$$LF = \int_{f=0.04}^{0.15} PSD(f)$$

  The PSD function represents the power spectral density and the variable f represents the frequency bands.

  While some studies consider LF as a reflection of sympathetic activity [Kim et al., 2018b], others associate it with both sympathetic and parasympathetic modulations [Shaffer and Ginsberg, 2017; Arakaki et al., 2023].

- **HF**: This feature represents the power intensity in the high-frequency band (0.15 Hz to 0.4 Hz) of the HRV signal and is mathematically represented as:

$$HF = \int_{f=0.15}^{0.4} PSD(f)$$

  The HF components indicate parasympathetic influence and a higher value indicates an increase in parasympathetic activity [Shaffer and Ginsberg, 2017; Kim et al., 2018b; Arakaki et al., 2023; Peabody et al., 2023].

- **SD1**: The poincaré plots (see Figire 2.16) are generated by mapping NN intervals and the previous NN intervals as ordered pairs [Claudia et al., 2003; Piskorski and Guzik, 2005]. The SD1 feature represents the dispersion of data points perpendicular to the identity line ($y = x$) in the Poincaré plot. In other words, it is the standard deviation of each point from the identity line and denotes the width of the ellipse [Piskorski and Guzik, 2005; Shaffer and Ginsberg, 2017].

$$SD1 = \sqrt{Var\left(\frac{NN_{i+1} - NN_i)}{\sqrt{2}}\right)}$$

  SD1 quantifies the short-term or beat-to-beat variability in the NN intervals and is similar to the RMSSD feature. So, an increase in SD1 indicates parasympathetic activation [Shaffer and Ginsberg, 2017].

- **SD2**: This feature quantifies the dispersion of data points along the identity line in the Poincaré plot. SD2 denotes the length of the ellipse and is calculated as the standard deviation of each point from the line $y = x + meanNN$ [Piskorski and Guzik, 2005; Shaffer and Ginsberg, 2017].

$$SD2 = \sqrt{Var\left(\frac{NN_{i+1} + NN_i)}{\sqrt{2}}\right)}$$

SD2 reflects the long-term or slower fluctuations in the NN intervals, which are influenced by both the sympathetic and parasympathetic branches of the ANS [Shaffer and Ginsberg, 2017].

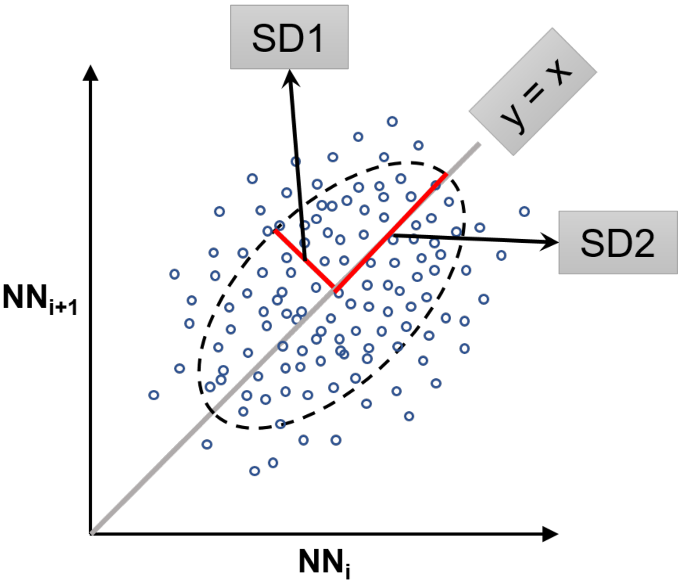


Figure 2.16: An illustration of a Poincaré plot between NN intervals (NN_i) and the subsequent NN intervals (NN_{i+1}). The identity line, SD1, and SD2 are labeled in the plot.

HRV during Stress

Stress is a high arousal state that increases sympathetic activity and reduces parasympathetic activity of ANS. This leads to an increase in heart rate, measured as the number of beats per minute. The increase in sympathetic activity during stress leads to a decrease in time-domain HRV features such as mean NN and SDNN. On the other hand, the reduced parasympathetic activity decreases RMSSD and pNN50 values [Purnamasari et al., 2019; Shaffer and Ginsberg, 2017; Kim et al., 2018b]. In the frequency domain, HF decreases, indicating reduced parasympathetic activity. The trend regarding LF is conflicting due to the decrease in parasympathetic activity and an increase in sympathetic activity.

Research has explored other HRV features that could be indicators of stress. A more extensive list of HRV features was utilized in Chapter 5 for stress detection. However, the relation between many of these features and ANS activations has not been investigated.

HRV during Flow

Considering the three-channel model, the three states lead to different ANS activations [Knierim et al., 2018; Peifer et al., 2014]. The anxiety state is similar to the stress state and is characterized by high arousal. It is associated with a decrease in HRV features, including mean NN, SDNN, RMSSD, pNN50, and HF, due to increased sympathetic activation and reduced parasympathetic activation. The boredom state is considered a low arousal state, with a lower heart rate and higher HRV. This state follows an opposite trend to stress in all the mentioned features.

The flow state is characterized by moderate arousal as well as relaxation. This leads to the activation of both sympathetic and parasympathetic branches of the ANS. While HRV features associated with sympathetic activation (e.g., mean NN) increase moderately, features associated with parasympathetic activation such as RMSSD and pNN50 also increase moderately. The flow state could lead to an increase in HF, reflecting the parasympathetic activity. Interestingly, the LF feature is expected to increase because of both sympathetic and parasympathetic activations.

2.3.3 Electrodermal Activity

Electrodermal activity (EDA), also known as galvanic skin response, is a signal that indicates the electrical conductance or resistance of the skin [Topoglu et al., 2020; Tronstad et al., 2022; Braithwaite et al., 2013; Posada-Quintero and Chon, 2020]. EDA is based on the principle that the skin's ability to conduct electricity varies with its moisture level, which is regulated by the sweat glands. Sweat is mostly made up of water and electrolytes, and thus, sweating leads to an increase in the electrical conductivity of the skin. The EDA sensors are typically placed on hands and feet due to the high density of sweat glands in these areas. However, it is important to note that EDA measurements are susceptible to various factors such as temperature and humidity.

The EDA signal contains two components: skin conductance level (SCL) and skin conductance response (SCR) [Topoglu et al., 2020; Tronstad et al., 2022; Braithwaite et al., 2013; Posada-Quintero and Chon, 2020]. SCL is the tonic component and represents the slow-moving part of the EDA signal. It indicates the baseline level of electrical conductivity in the skin. SCR is the phasic component and represents the fast-changing part of the EDA signal. It captures event-related changes in skin conductance that occur in response to specific stimuli. Due to the rapid changes, the SCR signal contains peaks characterized by sharp rises and slow decline.

Many techniques have been proposed to extract the SCL and SCR components from the EDA signal [Topoglu et al., 2020; Posada-Quintero and Chon, 2020]. A widely used one is the cvxEDA algorithm [Greco et al., 2015], which is shown to be robust against noise. This algorithm models the EDA signal as a sum of SCL, SCR, and noise. It applies a convex optimization approach to decompose the signal into the individual components.

EDA Features

EDA is a well-established marker of sympathetic activation of the ANS [Topoglu et al., 2020; Tronstad et al., 2022; Posada-Quintero and Chon, 2020]. Features are typically extracted from the raw EDA, SCL, and SCR signals. The widely used features from EDA and SCL signals are simple statistical metrics, whereas various peak-related features are extracted from SCR. This could be because many of the studies employ EDA in detecting stress or arousal, and SCR reflects the changes caused by the stimuli. Some of the popular EDA features are described below.

- **Mean and SD:** The simplest yet popular statistical features computed in many studies utilizing EDA are mean and standard deviation [Giannakakis et al., 2019; Horvers et al., 2021; Topoglu et al., 2020; Yu et al., 2018]. These features are computed for raw EDA, SCL, and SCR signals. The general formulas for computing these features are:

$$meanSig = \frac{1}{N}\sum Signal$$

$$SDSig = \sqrt{\frac{1}{N-1}\sum (Signal - meanSig)^2}$$

  Here, $Signal$ can be any of the three signals (raw EDA, SCL, or SCR) and N refers to the length of the signal. For a 60-second long EDA signal sampled at 50 Hz, the length of the signal is $60 \times 50 = 3000$.

- **Range**: This feature is typically computed for the raw EDA signal as $Range = max(EDA) - min(EDA)$. Some studies utilize the minimum and maximum values as features instead of computing the range [Horvers et al., 2021].

- **Number and Amplitude of SCR Peaks**: The features derived from the SCR peaks are widely used in many studies [Horvers et al., 2021]. The number of SCR peaks and the total amplitude of these peaks are frequently computed features. Figure 2.17 shows an example of a peak in the SCR signal.

- **Rise and Recovery times**: Rise time is the duration between the onset and the peak point. Similarly, recovery time is the duration between the peak point and offset. Figure 2.17 visualizes the rise and recovery times for a peak in the SCR signal. While the peak points are identified using a peak finding algorithm, the onset and offset points are identified based on threshold values [Horvers et al., 2021]. Features derived from these variables are also used in some studies targeting stress detection [Giannakakis et al., 2019].

EDA during Stress and Flow

The sweat glands are controlled by the sympathetic nervous system, which is responsible for the body's fight-or-flight response. When an individual experiences a high arousal

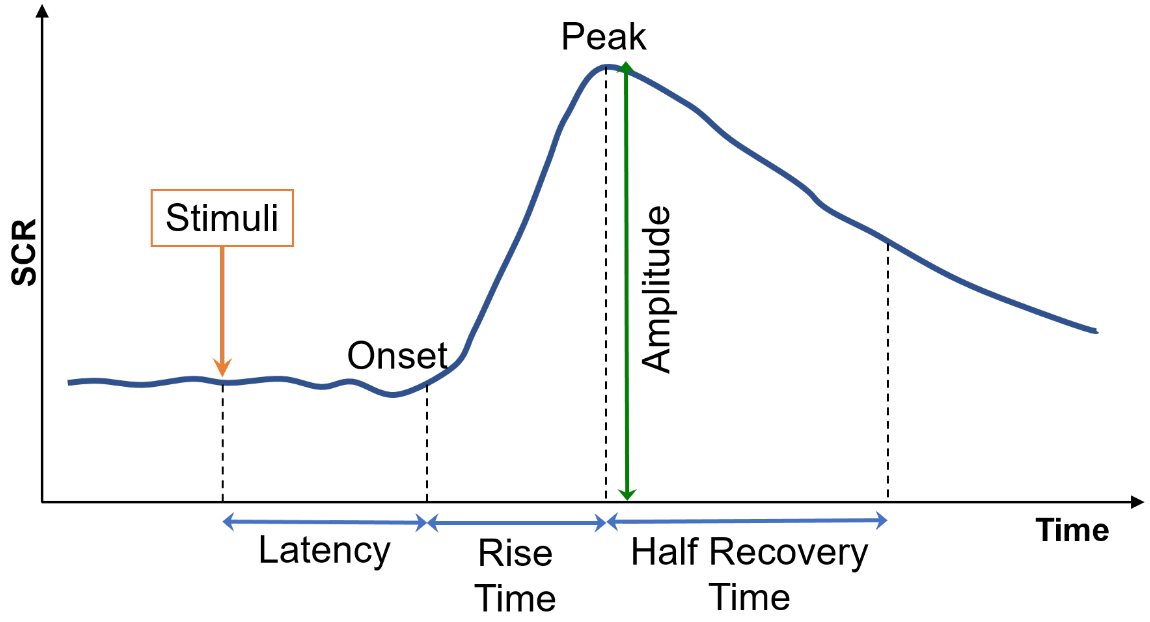


Figure 2.17: An illustrative plot of an SCR signal, labeled with onset, and peak points. The amplitude, rise time, and half recovery time are depicted in the plot.

state, such as stress or anxiety, the sympathetic nervous system activates, causing the individual to sweat more. The higher the amount of sweat, the higher the skin conductivity and the higher the EDA. Both SCL and SCR components increase with sympathetic activity. So, an increase in statistical features associated with these signals indicates sympathetic activity and arousal [Giannakakis et al., 2019]. Moreover, high-arousal stimuli also cause event-based changes in SCR, increasing the SCR peaks [Giannakakis et al., 2019].

While the EDA responses during stress are relatively known, it is not an efficient biomarker for differentiating various states in the flow model [Knierim et al., 2018]. A plausible reason is that the EDA is a signal governed almost exclusively by the sympathetic activity, without influence from the parasympathetic branch [Giannakakis et al., 2019; Topoglu et al., 2020; Tronstad et al., 2022; Posada-Quintero and Chon, 2020]. However, the flow state involves both sympathetic and parasympathetic activations. Although EDA could differentiate anxiety from boredom, it is not the most optimal for single-modal flow detection. However, the EDA signal is still a promising physiological signal in multimodal flow detection scenarios.

2.4 Machine Learning

Artificial Intelligence (AI) is defined as equipping machines with the ability to perform tasks typically associated with human intelligence, such as learning and problem-solving [Goodfellow et al., 2016]. Machine learning is a field of AI that focuses on developing algorithms and statistical models that learn an optimal representation of the data to recognize patterns or make predictions [Jordan and Mitchell, 2015; Goodfellow et al., 2016; Mahesh, 2020]. Although research has identified many techniques for machine learning, there are two categories of learning that are widely used: supervised and unsupervised learning [Bishop, 2006; Baştanlar and Özuysal, 2014; Nasteski, 2017; Mahesh, 2020; Janiesch et al., 2021]. Supervised learning involves training models using labeled data, where the input data is mapped to known output labels or values. The goal is to learn the mapping function for a given dataset. Typical supervised machine learning problems involve classification (e.g., predicting a discrete emotion label) and regression/estimation (e.g., predicting continuous values of valence-arousal) tasks. In unsupervised learning, the model is trained on unlabeled data, with the goal of discovering inherent patterns, statistical regularities, or relationships within the given dataset. Some common tasks under this learning paradigm include clustering (grouping similar data points) and dimensionality reduction (reducing the number of features or dimensions used to represent data while preserving the most relevant information). This thesis predominantly utilizes supervised machine learning models to predict various worker states.

2.4.1 Model Development

Phases

A machine learning model development typically follows three phases: training, validation, and testing [Baştanlar and Özuysal, 2014]. The training phase involves learning the mapping function that can predict the output label/value for a given input. This phase utilizes a training set, which comprises a majority of the data (typically around 80%) from the dataset. The validation phase utilizes a small subset of data, also called the validation set, to assess the performance of the trained model. The testing phase evaluates the performance of the model on unseen data, also called the testing set. The testing set can be an external dataset (different than the training dataset) or may originate from the same dataset as the training and validation sets (different data points from the same dataset).

Evaluation Methodology

K-fold cross-validation and leave-one-subject-out (LOSO) cross-validation are two common techniques used for training and evaluating the performance of machine learning models [Bishop, 2006; Baştanlar and Özuysal, 2014; Badillo et al., 2020]. K-fold cross-validation is a resampling technique that involves partitioning the available data into k equal-sized subsets or “folds". The model is trained using data from k-1 subsets and evaluated on the remaining subset. This process is repeated k times, where each iteration corresponds to one subset serving as the validation set. The final performance metric is calculated by averaging the results across all k iterations. LOSO cross-validation is a special case of cross-validation that is employed in scenarios where the data originates from a set of individuals. In LOSO, the model is trained on data from all subjects except one, and the left-out subject’s data is used for validation. This process is repeated, leaving out a different subject each time, until all subjects have been used for validation once.

LOSO evaluation is crucial in assessing the model’s ability to generalize to new, unseen subjects. K-fold cross-validation may not capture this aspect, as the folds can include data from the same subject in both the training and validation sets. This can lead to models achieving lower performance during LOSO evaluations, making it a stricter evaluation technique.

Performance Metrics

Performance metrics are quantitative measures used to assess the performance of machine learning models. These metrics are different for classification and estimation tasks. This thesis uses Accuracy and F1-score as performance metrics for classification tasks, whereas Concordance Correlation Coefficient (CCC) and Root Mean Squared Error (RMSE) for estimation tasks [Bajaj, 2023; Janiesch et al., 2021; Baştanlar and Özuysal, 2014; Badillo et al., 2020].

Both Accuracy and F1-score are based on the True classes (ground truth) and predicted classes (positive or negative). Accuracy measures the proportion of correct predictions made by the model out of all predictions and is given by:

$$Accuracy = \frac{\left(TruePositive + TrueNegative\right)}{\left(TruePositive + TrueNegative + FalsePositive + FalseNegative\right)}$$

$TruePositive$ is the number of positive class samples that the model predicted correctly, whereas $TrueNegative$ is the number of negative class samples predicted correctly. $FalsePositive$ is the number of negative class samples that were predicted incorrectly as the positive class. Similarly, $FalseNegative$ is the number of positive class samples that were predicted incorrectly as negative class.

For the positive class, precision measures the proportion of correct positive predictions out of all positive predictions made by the model. Recall measures the proportion of correct positive predictions out of all positive samples. The F1-score is a measure that combines precision and recall into a single metric. It is the harmonic mean of precision and recall, providing a balanced evaluation of a model's performance.

$$Precision = \frac{TruePositive}{\left(TruePositive + FalsePositive\right)}$$

$$Recall = \frac{TruePositive}{\left(TruePositive + FalseNegative\right)}$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{\left(Precision + Recall\right)}$$

Both Accuracy and F1-score metrics range between 0 and 1, with 1 indicating a perfect model that predicted all samples correctly. While accuracy is a commonly used metric, it can be misleading in cases where the dataset is imbalanced (i.e., one class is significantly more prevalent than the other). In such cases, the class-wise averaged F1-score may be more appropriate for evaluating the model's performance [Czakon, 2023].

For estimation tasks, RMSE measures the difference between the ground-truth value and the predicted value. On the other hand, CCC is a measure of similarity between a list of ground-truth values and predicted values. They can be computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(YTrue_i - YPred_i)^2}$$

In a list of N samples, $YTrue_i$ refers to the true value of the i^{th} sample and $YPred_i$ is the corresponding estimated value. A lower RMSE value indicates better model performance, as it implies that the predicted values are closer to the ground-truth values. The lowest plausible value is 0, indicating perfect agreement between predicted and true values.

$$CCC = \frac{2 \times CoVar(YTrue, YPred)}{\left(mean(YTrue) - mean(YPred)\right)^2 + SD(YTrue) + SD(YPred)}$$

Here, $CoVar(L1, L2)$ represents the covariance between two lists $L1$ and $L2$, $mean(L1)$ refers to the mean value of the list $L1$, and $SD(L1)$ is the standard deviation of the list $L1$. The CCC ranges from -1 to 1, with 1 indicating perfect agreement between the predicted and true values, 0 indicating no agreement, and -1 indicating perfect inverse agreement.

2.4.2 Shallow Models

Numerous machine learning algorithms have been proposed over the years [Bishop, 2006; Goodfellow et al., 2016; Janiesch et al., 2021]. They can be broadly classified as shallow and deep learning models. Shallow models are older machine learning techniques that follow simple algorithms or architectures, making them computationally efficient. As shown in Figure 2.18, they often rely on domain-specific knowledge and feature engineering, where relevant features are manually selected or hand-crafted from the raw data. The shallow models used in this thesis are described below.

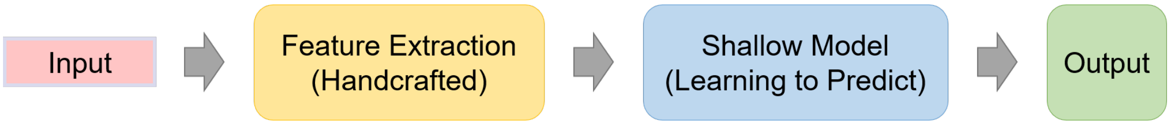


Figure 2.18: A typical shallow model pipeline.

Random Forest Classifier

A random forest classifier (RFC) [Breiman, 2001] is an ensemble learning method used for classification tasks. As illustrated in Figure 2.19, the main idea of RFC is combining multiple decision trees to form a “forest" [Badillo et al., 2020]. A decision tree is made up of nodes and edges [Nasteski, 2017; Mahesh, 2020]. It starts with the root node, which does not have an incoming edge. All other nodes have one incoming edge. The tree ends with a set of leaf nodes that represent the predictions of the decision tree model. The intermediate nodes or internal nodes represent decision points or rules that are discrete functions of input data.

The RFC algorithm creates subsets of the original training data through bootstrapping. That is, the samples in the subsets are randomly selected from the dataset with replacement, resulting in some samples being repeated while others are left out. For each subset, a decision tree is trained using the bootstrapped data. During the training phase of each tree, a random subset of features is selected at each node split rather than considering all features.

For each test sample, the RFC model aggregates the prediction of the individual trees through techniques such as averaging or majority votes. Although individual trees are prone to overfitting, the RFC model is considered robust as it mitigates this risk through ensemble learning [Jiang et al., 2020; Badillo et al., 2020].

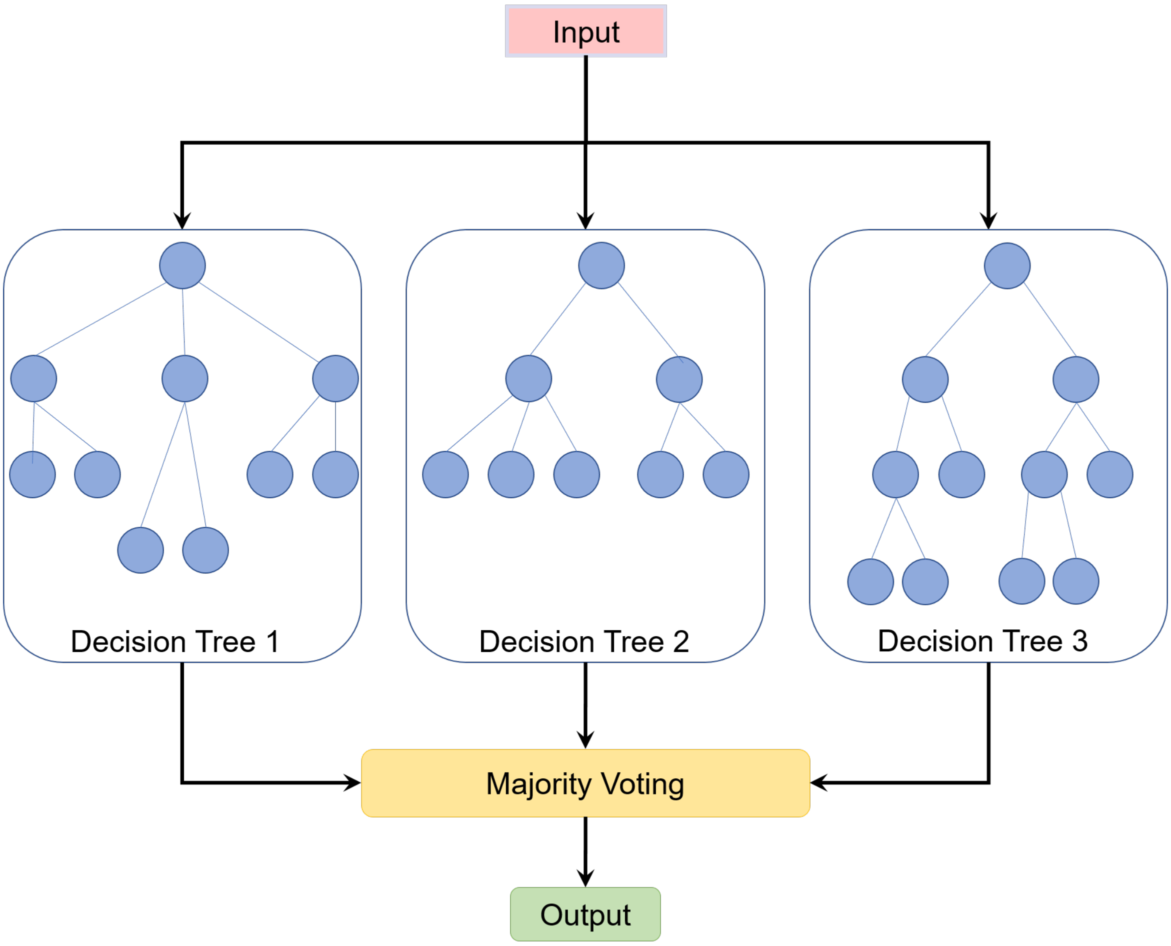


Figure 2.19: A visualization of an RFC model consisting of three decision trees with majority voting prediction aggregation.

Support Vector Machines

The goal of a support vector machine (SVM) [Cortes and Vapnik, 1995] model is to find the hyperplane that maximally separates the classes in the feature space. A hyperplane is a flat subspace that separates the data into classes. In two dimensions, the hyperplane is a line; in three dimensions, it is a plane, and so on. The hyperplane is chosen such that the distance between the closest data points (called the support vectors) and the hyperplane is maximized. This distance is known as the margin. For a binary classification of linearly separable data, the SVM is modeled as [Goodfellow et al., 2016; Baştanlar and Özuysal, 2014]:

$$decision\ boundary = w^T x + c$$

$$w = \sum_i \alpha_i y_i x_i$$

Here x_i represents a feature vector in the dataset, y_i is the corresponding label ($y_i \in \{-1, 1\}$), and α is a vector of coefficients. For any sample, if the $decision\ boundary$ equation is positive, then the label $y = 1$, and if the equation evaluates to negative, then $y = -1$.

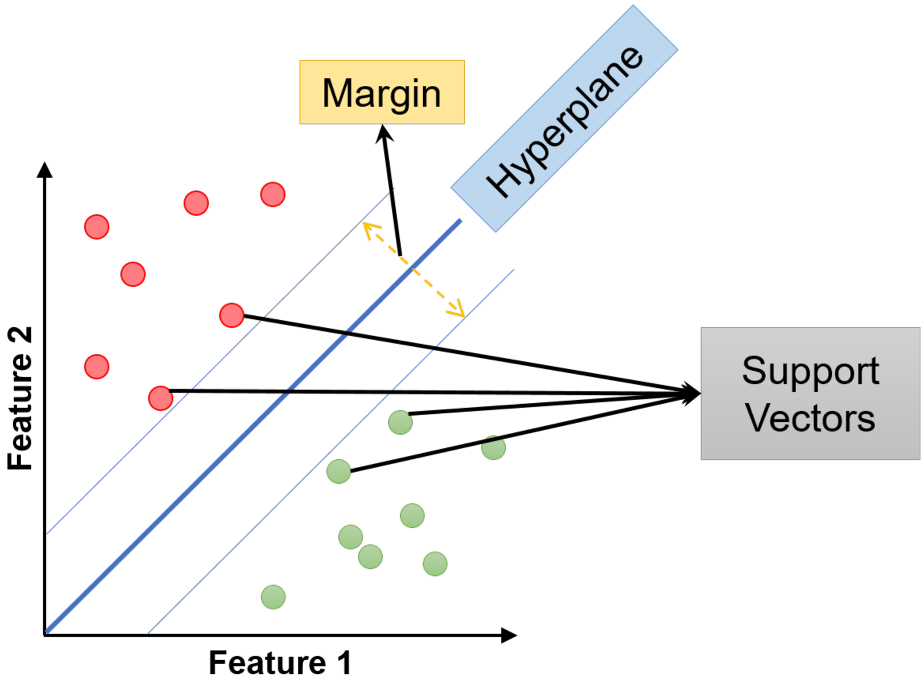


Figure 2.20: A visualization of an SVM model for linear separable data represented in terms of two features. The hyperplane (line in this case), margin, and support vectors are labeled in the image.

Kernel functions are utilized when the decision boundary is not linear [Goodfellow et al., 2016; Baştanlar and Özuysal, 2014; Badillo et al., 2020; Mahesh, 2020]. In this technique, the input data is mapped onto a high-dimensional feature space where a linear separation is possible. Subsequently, a hyperplane decision boundary is determined following the same procedure as linearly separable data. Radial basis function and linear kernels are commonly used in training SVMs.

To handle cases where the data is not perfectly separable, SVM utilizes a “soft margin" that tolerates some misclassification. This is controlled by a regularization parameter C that represents the trade-off between maximizing the margin and minimizing the classification error.

Simple Artificial Neural Networks

An artificial neural network (ANN) is a machine learning model that follows a graph-like structure made up of neurons or perceptrons [Bishop, 2006; Goodfellow et al., 2016; Badillo et al., 2020]. There are many neural network architectures that have been proposed in the literature. A neural network can be shallow or deep depending on the depth or number of layers in the architecture, which changes the model's learning capabilities [Janiesch et al., 2021]. The shallow models considered in this thesis are simple feedforward neural networks. A simple neural network involves an input layer, a few hidden layers, and an output layer. Neurons in each layer are connected to neurons in the subsequent layer through weighted connections. As visualized in Figure 2.21, each neuron in the hidden and output layers computes a weighted sum of its inputs from the previous layer and applies an activation function (e.g., sigmoid, ReLU) to produce its output. The output from the final layer represents the prediction of the model. The shallow networks in this thesis are called feedforward because information flows in a single direction, from the input layer through the hidden layers to the output layer, without any cycles or loops. Figure 2.22 presents an example representation of the simple ANNs used in this thesis.

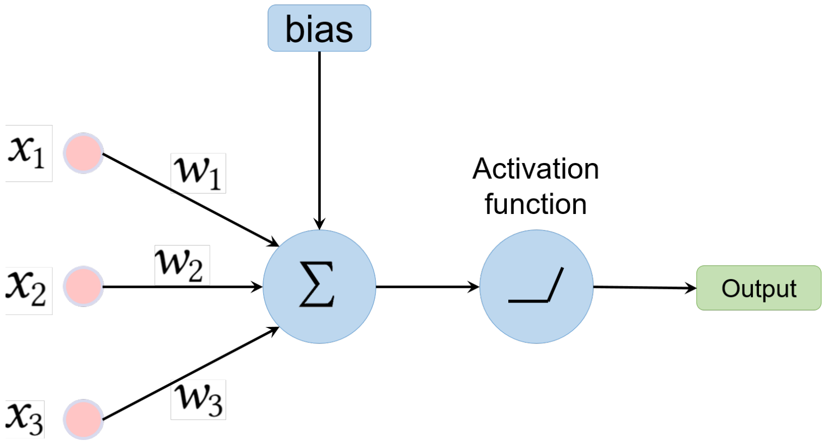


Figure 2.21: An illustration of operations associated with a simple neuron activation.

The training of a neural network involves three main components: loss function, backpropagation, and optimization algorithm. The loss function computes the loss for each input by calculating the difference between the predicted output and true values. Commonly used loss functions include mean squared error for regression tasks and cross-entropy loss for classification tasks. The errors (gradients of the loss function) are propagated backward through the network, starting from the output layer. The errors computed during backpropagation are used to update the weights and biases of the neural network using an optimization algorithm, such as stochastic gradient descent (SGD) and Adam. The optimization algorithm adjusts the weights and biases in the direction that minimizes the loss function, with the learning rate determining the step size. A higher learning rate leads to larger changes in weights. The training steps are repeated for a certain number of epochs or iterations until the neural network converges to a solution where the loss function is minimized.

2.4.3 Deep Learning Models

Deep learning models are neural networks that are “deeper" than simple feedforward models, i.e., they have many hidden layers. [Goodfellow et al., 2016; Janiesch et al., 2021]. In

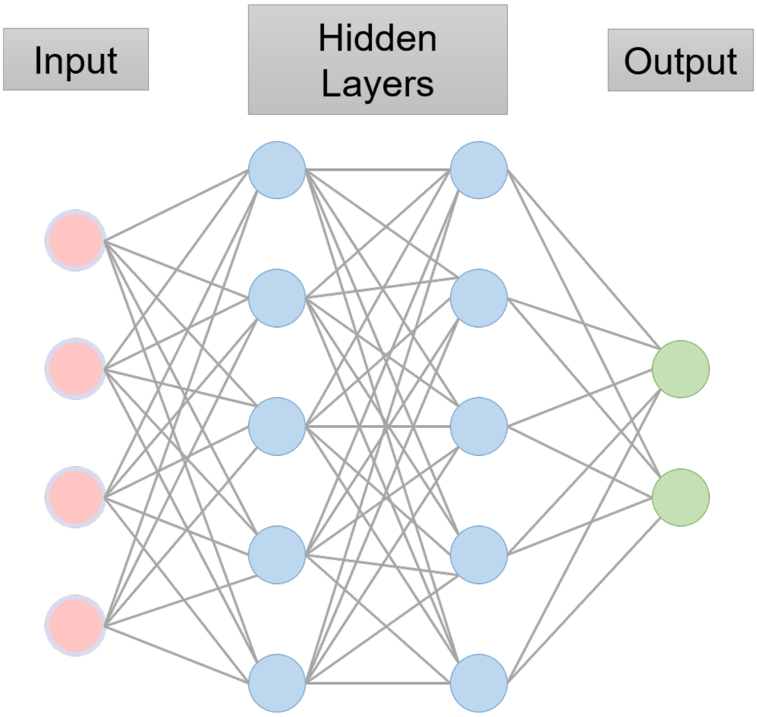


Figure 2.22: Visualization of a simple feedforward neural network with an input layer, two hidden layers, and an output layer.

recent years, neurons capable of advanced computations like convolution and feedback loops (recurrent connections) have been used in constructing convolutional neural networks (CNNs) and recurrent neural networks. Deep learning models are more complex and capable of learning data representations from the raw data instead of relying on handcrafted features. Consequently, these models are computationally heavy, and specialized hardware like GPUs are often employed to train them.

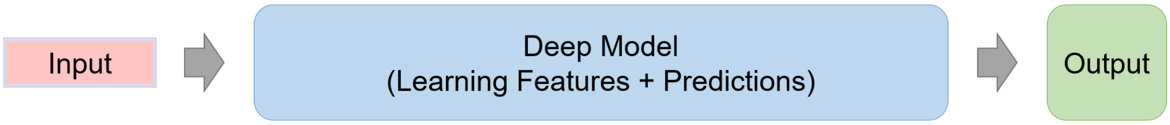


Figure 2.23: A typical deep learning model pipeline. Compared to shallow models in Figure 2.18, deep learning models learn the relevant features, eliminating the need for explicit feature extraction.

Convolutional Neural Networks

CNNs are deep learning models that are typically used for processing structured grid-like data (e.g., images). They are widely used in computer vision tasks such as image classification and object detection, as they can learn spatial features from input data. The work by LeCun et al. [1998] is considered the foundation of modern CNNs [Goodfellow et al., 2016; Gu et al., 2018].

The core component of CNNs is convolutional layers, which apply convolution operations to the input [Goodfellow et al., 2016; O'Shea and Nash, 2015; Gu et al., 2018; Dishar and Muhammed, 2023]. The convolution operation involves sliding a filter (or kernel) across the input, computing dot products between the filter weights and the input values to generate feature maps. An example of the convolution operation with a 2 × 2 filter is presented in Figure 2.24. Each filter detects specific local patterns such as edges, textures, or more complex features in deeper layers. These filters are learned during the training process.

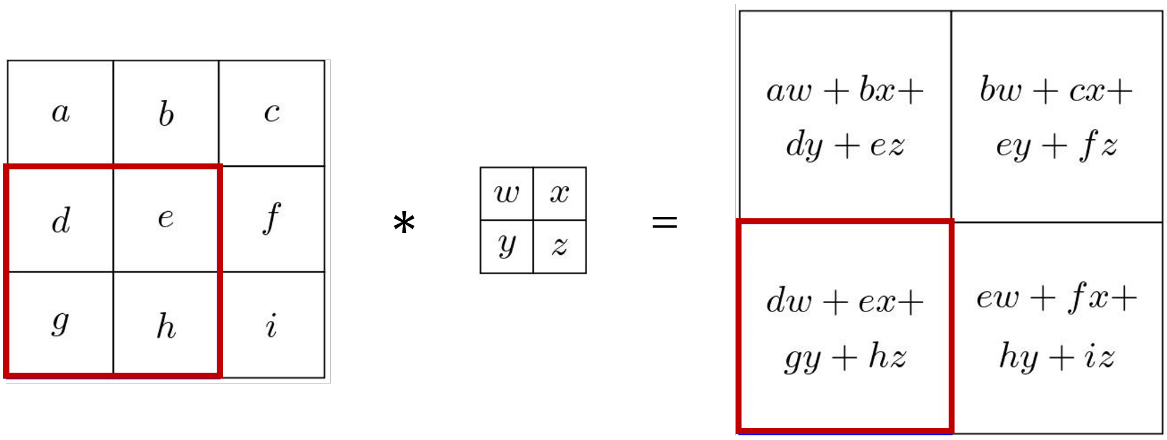


Figure 2.24: An example of convolution operation for a 3 × 3 input and a 2 × 2 filter.

The pooling layer is another widely used layer in CNNs [Goodfellow et al., 2016; O'Shea and Nash, 2015; Gu et al., 2018]. The pooling function replaces values at a location by summary statistics (e.g., average, maximum) of nearby values. It can be seen as a downsampling layer, which reduces the dimensions of the feature maps. After several convolutional and pooling layers, the feature maps are flattened and fed into one or more fully connected layers. These layers are similar to the hidden layers of a simple ANN. They utilize the features extracted through convolutional layers to generate predictions.

Batch normalization and dropout layers are often used in CNN architectures to reduce overfitting [Gu et al., 2018; Dishar and Muhammed, 2023]. During training, for each mini-batch, the batch normalization layer applies z-normalization (subtracting the mini-batch mean and dividing by the mini-batch standard deviation) to the input. This layer stabilizes the learning process by reducing the dependence on the initial values of the weights. The dropout layer randomly drops a fraction (set as a hyperparameter) of the neuron activations in a layer during the training phase. This prevents any single neuron or feature from having an excessive influence on the output.

VGG-16 Architecture

Throughout this thesis, the VGG-16 architecture is utilized as a basis for the image-based deep learning models. This architecture was proposed by Simonyan and Zisserman [2014], and is named after their group (Visual Geometry Group). The original model proposed for object recognition tasks had 16 weighted layers (13 convolutional layers and 3 fully connected layers).

The model takes as input RGB images (three channels), typically scaled to 224 × 224 pixels (default dimensions). The base network (without fully connected layers and output layer) is made up of five convolutional blocks, stacked one after the other. Each block contains two or three convolutional layers, followed by a max pooling layer. All convolutional layers utilize filters of size 3 × 3. The number of filters increases with each block, starting with 64 filters in the convolutional layers of the first block and doubling after each block, reaching 512 filters in the last block.

The final layers of the model are modified following the approach proposed by Lin et al. [2014]. The last pooling layer is replaced with a global average pooling over all feature maps. After this layer, a fully connected layer with 1024 neurons is added, which in turn is connected to the output layer. Figure 2.25 visualizes the VGG-16 architecture followed in this thesis.

Transfer Learning

Transfer learning is a machine learning technique that involves transferring knowledge gained from one task or domain to another related task or domain. This technique is particularly useful when the target task has limited training data. There are many classifications of transfer learning [Zhuang et al., 2020]. One classification distinguishes between inductive, transductive, and unsupervised transfer learning [Pan and Yang, 2009]. In inductive transfer learning, the source and target domains are the same, but the tasks are different but related. Transductive transfer learning involves the same source and target tasks, but the domains are different but related. Unsupervised transfer learning involves unsupervised learning tasks (e.g., clustering, dimensionality reduction) with both source and target domains containing unlabeled data. This thesis focuses on inductive transfer learning, where the learned feature representations are transferred.

As highlighted by Yosinski et al. [2014], there are two primary methods for transfer learning feature representations in deep neural networks: freezing and fine-tuning. In the freezing method, weights from the initial layers of a pre-trained model are copied to the target model, and these layers are designated as “frozen" or not trainable. Meanwhile, the remaining layers, marked “unfrozen", are initialized randomly and trained on the target dataset. The fine-tuning method also involves transferring weights from the initial layers of a pre-trained model and initializing the remaining layers with random weights. However, in fine-tuning, no layer is marked as frozen; instead, all layers undergo further training on the target dataset.

2.5 Background Summary

This chapter presented the psychological backgrounds of relevant user states, the behavioral and physiological signals that are indicators of these states, and the various machine learning models utilized in this thesis. Various theories regarding the manifestation of emotion, pain, stress, and flow were briefly described in the psychological concepts section. Discrete emotion models like Ekman’s basic emotions and continuous emotion models like the pleasure-arousal-dominance models were discussed. Theories about the pain experience and how it can be modulated by emotions and stress were also discussed. The concept of stress was explored from the perspective of stressors (stimuli), stress response,

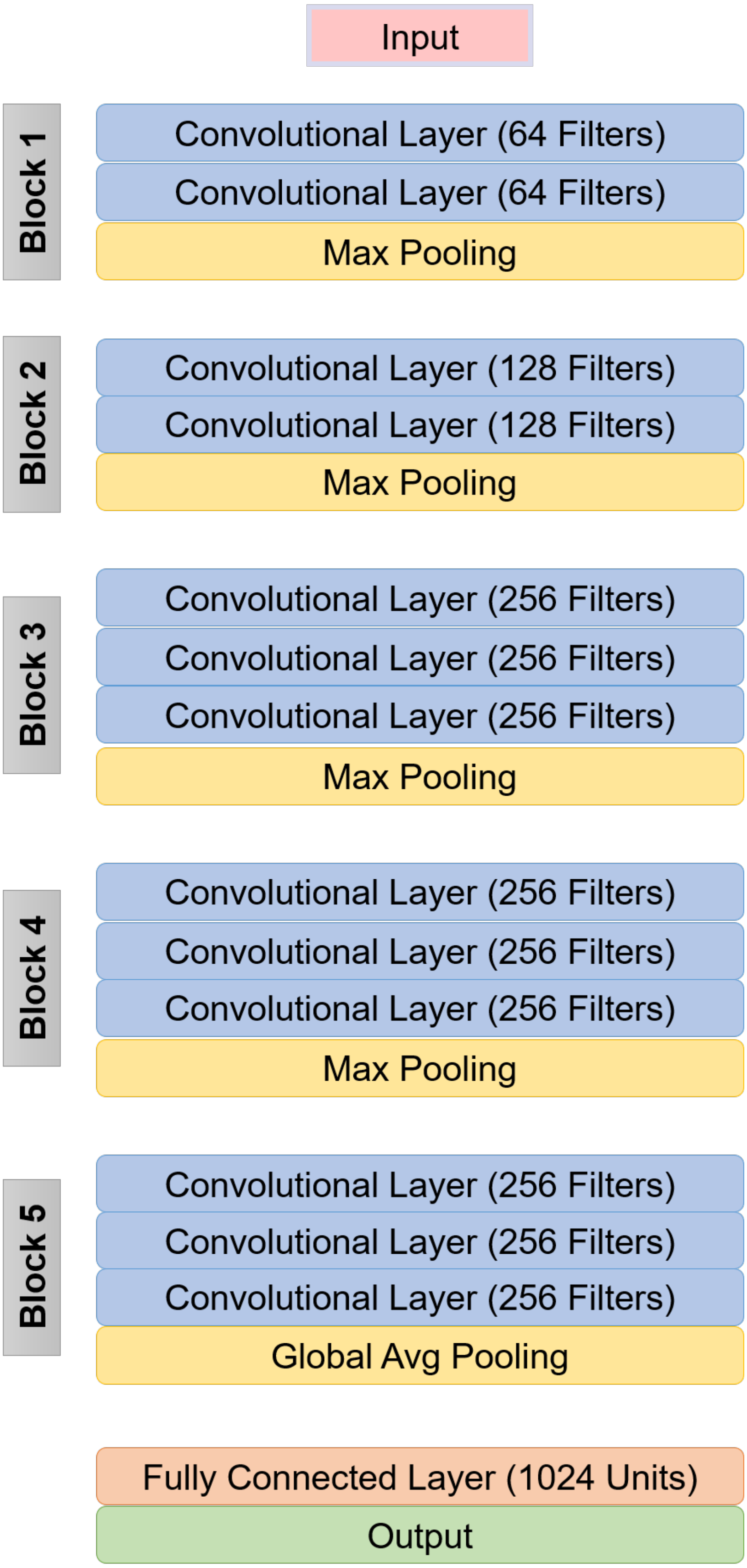

Input
Block 1
Convolutional Layer (64 Filters)
Convolutional Layer (64 Filters)
Max Pooling
Block 2
Convolutional Layer (128 Filters)
Convolutional Layer (128 Filters)
Max Pooling
Block 3
Convolutional Layer (256 Filters)
Convolutional Layer (256 Filters)
Convolutional Layer (256 Filters)
Max Pooling
Block 4
Convolutional Layer (256 Filters)
Convolutional Layer (256 Filters)
Convolutional Layer (256 Filters)
Max Pooling
Block 5
Convolutional Layer (256 Filters)
Convolutional Layer (256 Filters)
Convolutional Layer (256 Filters)
Global Avg Pooling
Fully Connected Layer (1024 Units)
Output

Figure 2.25: An illustration of VGG-16 architecture utilized in this thesis for image-based predictions.

and cognitive evaluation of the situation (challenges and resources). Finally, the modeling of flow state and related experiences (boredom, anxiety) in terms of challenge and skill were presented.

The next sections focused on the various affective and social signals. Facial behavioral signals including facial expressions and gaze were discussed, along with their potential in emotion, pain, and visual attention recognition. The section on physiological signals provided a brief overview of the role of ANS in regulating physiological responses. Two signals - HRV and EDA - were described including the signal characteristics, various features, and how they vary depending on the different ANS activations. A brief description of how these signals reflect stress and flow states was also presented.

The last part of this chapter focused on the machine learning concepts. The phases of model development and the techniques used in these phases were discussed. The concepts of shallow and deep learning models were presented, along with some architectures for each type of learning. In the subsequent chapters, the signals and features are used to train various models for detecting the relevant states.

Part II

Transferability, Generalizability, and Replicability

Chapter 3

Attention and Distraction Recognition

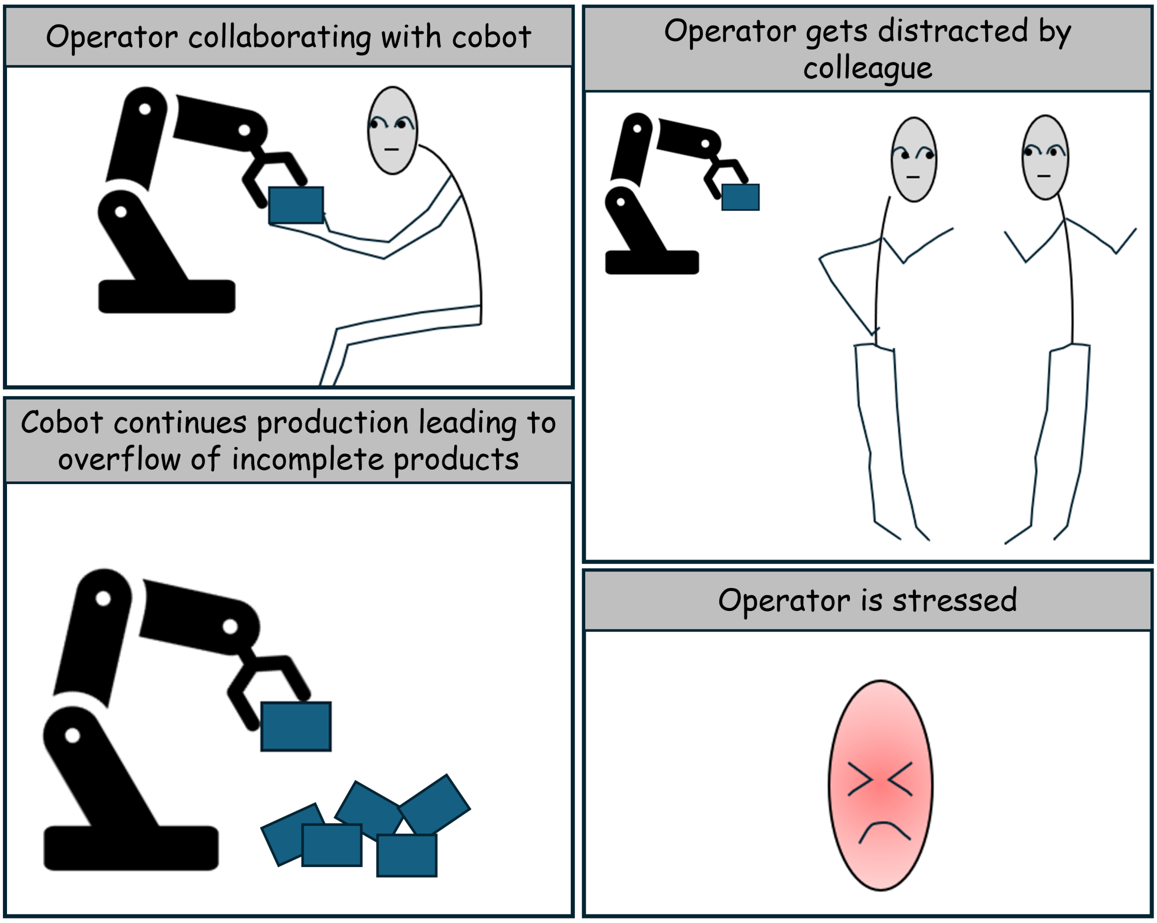

Operator collaborating with cobot
Operator gets distracted by colleague
Cobot continues production leading to overflow of incomplete products
Operator is stressed

Figure 3.1: A comic strip illustration of an industrial Human-Robot Collaboration (HRC) use-case where attention/distraction detection can mitigate negative experiences. In this situation, a worker collaborating with a cobot on a production line becomes distracted. The non-adaptive cobot continues production, resulting in excess production of incomplete items, ultimately leading to a stressful situation when the worker returns.

3.1 Overview

In Industry 5.0 scenarios, the gaze of a worker provides valuable information that a cobot can utilize to adapt its behavior. Research in this domain often focuses on adaptations that enhance productivity and safety, while the aspect of worker well-being is relatively unexplored [Nicora et al., 2021; Fan et al., 2022]. Imagine a scenario depicted in Figure 3.1, where a brief lapse in the worker's attention results in a highly stressful circumstance. In this case, the cobot can either slow down its production rate or place the assemblies in a dedicated location, and thus avoid disruption of the assembly line. Such human-centered adaptations can reduce negative experiences and psychological stress on the worker.

This work utilizes the direction of gaze as an indicator of an individual's attention, effectively conveying their current area of interest. The primary objective of previous related studies [Saran et al., 2018; Shi et al., 2021; Huang and Mutlu, 2016] was identifying objects on a table that the participant is currently observing. Drawing inspiration from the field of driver attention detection [Ahlstrom et al., 2013; Vora et al., 2017; Tayibnapis et al., 2018], this chapter proposes a gaze-based model to categorize operator's attention and distraction. The underlying concept of developing this model involves a two-step process: the initial training of a model to detect gaze direction, followed by mapping this detected direction to predefined classes of attention. The model utilizes images captured from a frontal camera as its input.

The model described in this chapter is used in Chapter 8 to analyze the operator's gaze behavior and subsequently adapt the cobot behavior. The contents of this chapter expand the research presented in:

- P. Prajod, M. Lavit Nicora, M. Malosio, and E. André. Gaze-based attention recognition for human-robot collaboration. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 140–147, 2023a

  [*I curated the datasets, including data processing and labeling. I also developed the machine learning models and performed the analysis. Furthermore, I contributed to the design of data collection setup.*]

- R. Arora, P. Prajod, M. L. Nicora, D. Panzeri, G. Tauro, R. Vertechy, M. Malosio, E. André, and P. Gebhard. Socially interactive agents for robotic neurorehabilitation training: Conceptualization and proof-of-concept study. *arXiv preprint arXiv:2406.12035*, 2024

  [*I contributed significantly to defining the framework and designing the data collection setup. Additionally, I developed the machine learning models and the real-time data processing and detection pipelines.*]

3.2 Background Literature and Previous Works

3.2.1 Distraction Datasets

Worker distraction detection is critical in industrial settings as it can lead to increased mistakes and faulty products. Gaze-based distraction detection has been studied thoroughly in applications like driver assistance and education, but it is not widely studied in the context of worker distraction. This section discusses some of the existing distraction datasets from various domains, tabulated in Table 3.1.

Paper	Context	Participant Data
Montoya et al. [2016] (SFDDD)	Driver	Face/Upper body images
Taylor et al. [2015] (Warwick-JLR)	Driver	ECG, EDA
Taamneh et al. [2017]	Driver	Face/Upper body videos, Eye tracker, EDA, Heart rate, Breathing rate
Abouelnaga et al. [2017] (AUCDD)	Driver	Face/Upper body images
Billah et al. [2018] (EBDD)	Driver	Face/Upper body videos
Jegham et al. [2020] (3MDAD)	Driver	Face/Upper body videos
Saad et al. [2020]	Driver	Face/Upper body images
Delgado et al. [2021]	Student	Face images
Asish et al. [2021]	Student	Eye tracker
Das et al. [2022]	Driver	Face/Upper body videos, Temperature, EDA, Heart rate, Breathing rate
Rahman et al. [2023] (SynDD1)	Driver	Face/Upper body videos
Shaiqur Rahman et al. [2022] (SynDD2)	Driver	Face/Upper body videos
Wang et al. [2023] (100-Driver)	Driver	Face/Upper body videos
Zaparas et al. [2023] (ARDIST)	Worker	Eye tracker
Dai et al. [2023]	Worker	Face/Upper body videos, Eye tracker
Kaewkaisorn et al. [2024] (RLDD)	Student	Face images

Table 3.1: A brief overview of existing human distraction datasets.

Distraction detection is widely studied in the context of driving. One of the most popular driver-distraction datasets is the State Farm Distracted Driver Dataset (SFDDD) [Montoya et al., 2016], which was published as part of a Kaggle competition. This dataset contains RGB images recorded using a dashboard camera. The SFDDD labeled nine classes of distraction, such as texting, drinking, and talking to passengers. Similar distraction activity labels were leveraged by other driver distraction datasets (e.g., Abouelnaga et al. [2017];

Billah et al. [2018]; Saad et al. [2020]; Rahman et al. [2023]). It is important to note that these labels are chosen based on frequent distracted driver behaviors and are not universal indicators of distraction. For example, a worker talking to the cobot does not suggest that the worker is distracted.

In naturalistic situations, driving can occur in the daytime or at night. The typical RGB images/videos that are reliable in daytime illumination may not be effective at night in low-lighting conditions. So, datasets such as 3MDAD [Jegham et al., 2020], SynDD1 [Rahman et al., 2023], and 100-Driver [Wang et al., 2023] additionally collected infrared camera recordings for effective low-lighting distraction detection. Both RBG and infrared cameras are typically placed near the dashboard, recording a side profile of the driver along with their upper body. These images/videos capture information about gaze as well as posture (e.g., hands off the wheel, holding a phone). In addition to distraction instances, SynDD1 [Rahman et al., 2023] and SynDD2 [Shaiqur Rahman et al., 2022] datasets contain samples of drivers looking at various pre-determined gaze zones of the car. These samples can be used to train models that predict the visual attention of the driver. A few datasets [Taylor et al., 2015; Taamneh et al., 2017; Das et al., 2022] collected physiological data to assess driver distraction.

Some datasets were collected to detect distraction in students in learning environments. While Asish et al. [2021] utilized gaze tracking in a VR environment to detect distraction, works like Delgado et al. [2021] and Kaewkaisorn et al. [2024] relied on dedicated facial images/videos. The cameras capturing RGB facial images are typically placed in front of the student. Unlike driver distraction datasets, the attention/distraction labels in these datasets are assigned based on the student's gaze direction (e.g., whiteboard, table/notebook, random direction) instead of specific behaviors.

Recent research has acknowledged the lack of worker distraction datasets, which led to the emergence of datasets designed for recognizing worker distraction in industrial contexts. Notable examples include the datasets proposed by Zaparas et al. [2023] and Dai et al. [2023]. The dataset collected by Zaparas et al. [2023] contains gaze-tracking data from participants working with a cobot in augmented reality settings. They incorporated visual (e.g., random objects) and audio (e.g., ambulance siren, honking) distracting stimuli. On the other hand, Dai et al. [2023] collected both frontal camera and wearable eye-tracker data during physical HRC sessions. They induced distractions by sudden clapping noise, a co-worker entering the workspace, and personal phones. However, this dataset is currently a pilot version that is very limited in size.

3.2.2 Gaze Detection from Facial Images in HRC

As seen from Table 3.1, facial/upper body images are the most popular modality in detecting distraction. The typical features extracted from facial images include gaze direction, facial expressions, and head pose. As mentioned in Chapter 2, the literature has shown that gaze is a prominent indicator of human attention and distraction. This section focuses on gaze-based attention/distraction in the HRC context. Many works in HRC settings [Huang and Mutlu, 2016; Newman et al., 2020; Shi et al., 2021; Chadalavada et al., 2018; Paletta et al., 2019; Chan et al., 2022; Gomez Cubero and Rehm, 2021] utilize wearable eye trackers to detect the gaze direction and in turn, the area that attracted the user's attention. Most of these works are described further in Chapter 8. In these works, the attention information is used in recognizing intent, which is based on the observation that humans typically direct their gaze towards an object before interacting with or manipulating it. However, wearable devices can be intrusive and less suitable for industrial settings since they require workers to wear extra equipment, thereby introducing additional prerequisites and protocols. For instance, workers may need training on how to wear the gaze tracking device correctly. As an alternative, non-intrusive solutions utilizing facial images/videos have gained traction. Hence, this section further narrows its focus by considering works that utilize facial images/videos to detect gaze-based attention/distraction.

There are mainly two ways to recognize gaze direction from a camera image - using eye region information and head pose. Palinko et al. [2016] compared these two methods during an HRC setting. The features for both methods were derived from frontal face images. These images were captured using the robot's (iCub robot) camera situated near its eye. They used head orientation to calculate head-gaze direction, while eye gaze was determined using information from the eye region, which in turn was computed from facial landmarks. To evaluate these methods, the researchers designed a collaborative puzzle experiment. Operators were tasked with assembling a tower using four blocks, which were held by an experimenter and the robot. Success in retrieving the blocks was contingent on the operators exhibiting specific gaze behaviors, involving looking at the blocks and then looking at the robot's face, or vice versa. Compared to head-gaze method, eye-gaze method led to more successful collaborations. Although the study's primary goal differed from typical intention recognition, the experimental design naturally inclined towards discerning gaze directed toward specific objects in space, namely two blocks and a stationary robot face.

Dufour et al. [2020] proposed using the visibility of the end-effector as a kinematic constraint for robot motion. This idea was inspired by literature suggesting that a robot should always perform its tasks within the operator's field of view to ensure the operator's safety and comfort. To implement this, the authors computed the participant's gaze direction from camera images by taking into account the positions of the shoulders, neck, and head. The researchers utilized an RGB-D camera facing the participant, mounted on the wall behind a Baxter robot. Initial tests of their system showed promise for automatically adapting the robot configuration based on the participant's line of sight.

Wong et al. [2023] developed a prototype for safety in HRC through the integration of camera and tactile sensors. The primary aim of their study was to differentiate between intentional and unintentional touches during interactions. They deployed an RGB-D camera positioned above the Baxter robot's head to capture color and depth data. For the extraction of relevant information, they employed OpenPose to identify body keypoints from the camera images. Gaze direction was subsequently computed based on these head keypoints. Using the features derived from gaze, body pose, and tactile sensor data, they trained shallow machine learning models to determine whether an operator's touch on the robot was intentional or not.

The studies discussed until now have very distinct areas of interest, making gaze direction estimation based on head pose reasonably effective. However, Saran et al. [2018]

proposed the use of frontal face images to discern the object that the participant is looking at (intention recognition), without relying on eye trackers. Such a task requires more precise estimation of gaze direction. They developed a deep-learning model to track a participant's gaze from the perspective of a robot. Their experimental setup involved a robot equipped with a Kinect camera positioned on its head, situated across a table from a participant. To train and evaluate their model, they collected an image dataset where each participant was asked to gaze at particular objects or the robot itself for a fixed duration.

Paper	Objective	Cross-dataset
Palinko et al. [2016]	Comparing gaze estimation methods	No
Saran et al. [2018]	Intention recognition	No
Dufour et al. [2020]	Robot within user's line of sight	No
Wong et al. [2023]	Intentional touch recognition	No
Prajod et al. [2023a]*	**Distraction detection**	**Yes**

*Table 3.2: An overview of the literature on attention recognition in HRC settings. The entry marked with * is expanded in the subsequent sections of this chapter.*

While the above studies may not have gaze estimation as their primary objective, it remains a significant underlying factor for precision-related issues. Hence, gaze estimation would be a good starting point for training models discerning attention and distraction.

3.2.3 Research Gap

As highlighted through Table 3.2, two significant research gaps have come to light, and these gaps will be the focal points of upcoming sections of this chapter: the lack of distraction detection and the absence of cross-dataset evaluations.

- **Distraction detection**: Current research in industrial HRC primarily emphasizes efficiency and safety through intent recognition. However, there is a notable gap concerning human factors, particularly in addressing the automatic detection of worker distraction. The lack of worker distraction datasets underscores this gap. Moreover, the two identified datasets were recently published, which implies that there is limited research on detecting worker distraction.

- **Cross-dataset validation**: The effectiveness of attention recognition is inherently tied to the specific experimental setup, making it difficult to apply these models to different settings. None of the studies discussed in Section 3.2.2 closely mimicked real-world industrial environments, and thus the applicability of their models to actual industrial scenarios is uncertain. Hence, it is imperative to validate such models in an industry-like use case. Moreover, facial image models including gaze estimation models need to be evaluated under various conditions, including diverse participants and recording scenarios, even though the setup layout is consistent between experiments.

3.3 From Gaze Estimation to Attention Recognition – A Transfer Learning Approach

While gaze-based distraction detection is relatively uncommon in the context of HRC, it is a prevalent topic in driver assistance systems. Driver attention and distraction are typically assessed by monitoring gaze zones within the vehicle. Research in this domain often relies on datasets captured through frontal cameras with participants gazing at predefined zones of interest within the car. Although the challenges and settings for driver distraction differ from those in HRC, the methodologies developed for addressing driver distraction could offer valuable insights for attention (and distraction) recognition within HRC scenarios.

Research by Vora et al. [2017] demonstrated the potential of deep learning models in detecting driver distraction, although these models typically demand extensive data. To circumvent this limitation, Tayibnapis et al. [2018] introduced a transfer learning approach. They connected a pre-trained neural network to an SVM, enabling the classification of driver's face images into gaze zones. Drawing inspiration from these studies, a transfer learning approach is taken to train a deep learning model for gaze-based attention recognition.

As mentioned earlier in Section 3.2.2, gaze estimation plays a pivotal role in the attention/distraction recognition process. The effectiveness of the attention model is greatly contingent on the precision of the underlying gaze estimation. So, in the proposed approach, a deep learning model is initially trained using a state-of-the-art gaze estimation dataset. Subsequently, a transfer learning technique is employed to establish a mapping between gaze directions and distinct attention classes. This approach is demonstrated using two use cases. The first one involves a straightforward scenario aimed at identifying whether the participant is focused on the screen or is distracted. In the second use case, which resembles an industrial assembly setup, the goal is to discern instances where the operator is engaged in assembly, directing their gaze towards the robot, or distracted.

3.3.1 Gaze Estimation Model

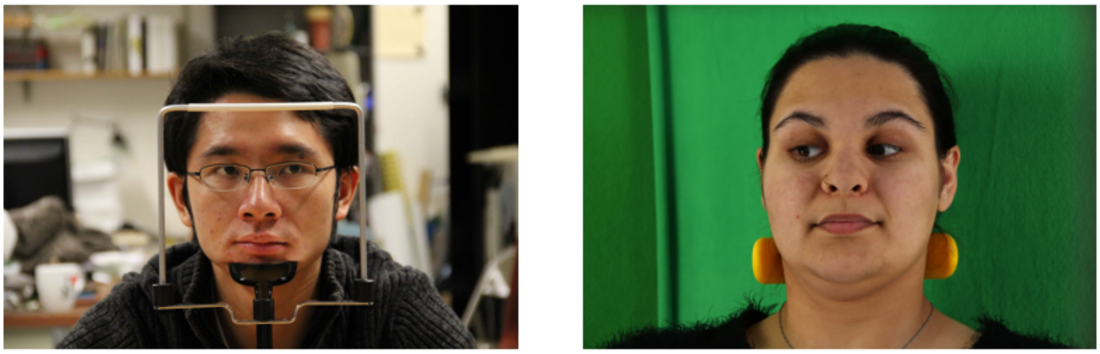


Figure 3.2: Illustrative images from the Columbia [Smith et al., 2013] (on the left) and ETH-XGaze Zhang et al. [2020] (on the right) datasets. These images depict participants with a forward-facing head pose while gazing towards the sides. The copyright remains with the respective dataset creators.

This section outlines the training process for the gaze estimation model, which forms the basis for the attention recognition model. Unlike several prior studies (refer to Section 3.2.2), the gaze direction is estimated based on eye gaze rather than head orientation. This decision was motivated by the goal of achieving a more precise gaze estimation model. While head orientation can serve as a reasonable proxy for gaze direction, disparities between the two can arise, as depicted in Figure 3.2.

ETH-XGaze Dataset

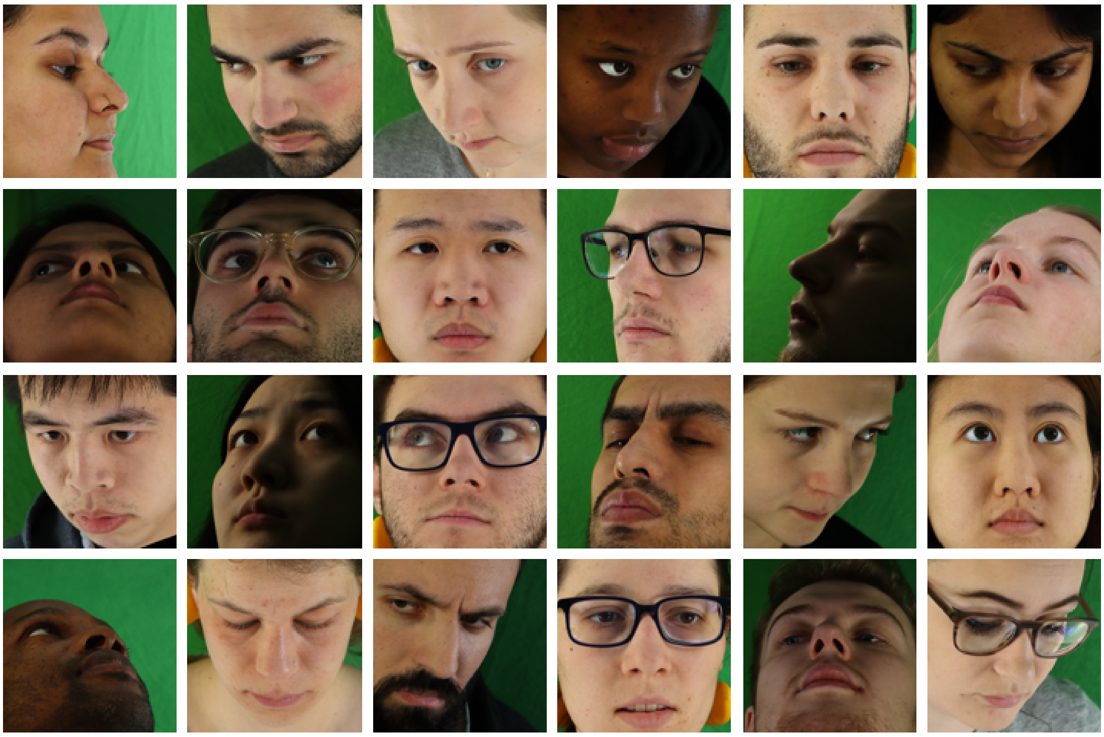


Figure 3.3: Sample images from the ETH-XGaze dataset illustrating variations in gaze direction, head pose, illumination, and other factors. The copyright remains with the dataset creators [Zhang et al., 2020].

The gaze estimation model is trained using the ETH-XGaze dataset Zhang et al. [2020]. The ETH-XGaze dataset is a large collection of high-resolution images designed for the training of robust gaze estimation models. It encompasses over one million images obtained from 110 participants with diverse attributes, including gender, age, and ethnicity. As shown in Figure 3.3, this dataset introduces variations in gaze, such as extreme angles and different head poses, under varying illumination conditions. The dataset labels these images with two continuous values, indicating the pitch and yaw of the gaze direction vector from the camera's perspective. Ground-truth labels for 80 participants are available for training models.

Training Procedure

The neural network architecture used in this task is VGG16 Simonyan and Zisserman [2014], comprising five convolutional blocks and pre-trained on the ImageNet Russakovsky et al. [2015] dataset. As illustrated in Figure 3.4, VGG16 is connected to a fully connected layer, and subsequently to a prediction layer for estimating gaze direction. Input images undergo preprocessing, involving face-cropping and scaling to conform to the standard VGG16 input dimension of 224 × 224. Among the labeled dataset consisting of data from 80 participants, 10% (8 participants chosen randomly) are reserved for validation, while the remaining data is employed for training. During training, images are processed in batches of 32. The mean absolute error is employed as the loss function. The Stochastic Gradient Descent (SGD) optimizer is used with an initial learning rate of 0.001, and the learning rate is reduced by a factor of 0.1 if the validation loss remains stagnant for five epochs. An early-stopping mechanism is implemented to prevent overfitting, whereby training stops if the validation loss doesn't decrease over the last 7 epochs.

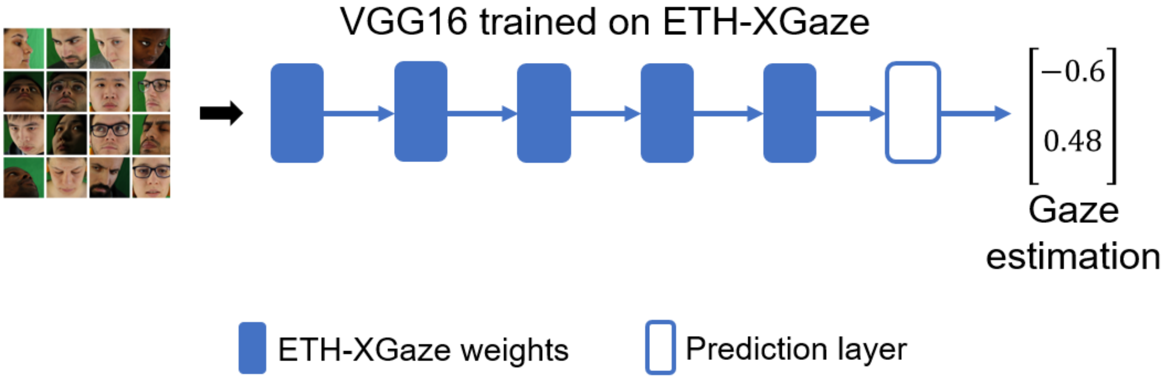

VGG16 trained on ETH-XGaze
$\begin{bmatrix} -0.6 \\ 0.48 \end{bmatrix}$
Gaze estimation
ETH-XGaze weights
Prediction layer

Figure 3.4: A depiction of the VGG16-based gaze estimation network. The boxes filled with blue color represent the convolutional blocks and the last outline box represents the fully connected and prediction layers. Gaze is estimated in terms of pitch and yaw

3.3.2 Use Case 1 – Screen vs. Distracted

In this initial use case, the primary objective is to determine if the participant is distracted, which is defined as the participant shifting their attention from the screen to their surroundings. Notably, this scenario does not involve human-robot collaboration. Instead, its focus is on assessing the feasibility of training a deep learning model that can effectively map the participant's gaze to areas of interest using a small dataset.

In the driver distraction scenario, the predefined gaze zones within the car provide a reference for determining the driver's attention. In contrast, gaze-based attention recognition in HRC is inherently tied to the specific characteristics of the setup, necessitating customized models. Acquiring large datasets suitable for training deep learning models in an industrial setting can be challenging. Therefore, the core question here is whether the estimated gaze direction can be accurately mapped to areas of interest with a limited number of samples.

Setup and Dataset

The experimental setup was relatively straightforward, involving participants positioned in front of a computer screen while following specific guided gaze instructions on where to look. To ensure the robustness of the dataset, participants were encouraged to vary their gazing angles and head poses during the data collection process. A total of 5 participants (3 males, 2 females; age: 24 to 52 years) took part in the study, with 3 of them wearing glasses. Each participant contributed approximately 20 images for each of the two predefined classes, namely “screen" and “distracted". Figure 3.5 shows some example images from the two classes. High-resolution 1920 × 1080 images of participants’ faces and gaze were captured using a Logitech C920 camera positioned on top of the computer screen.

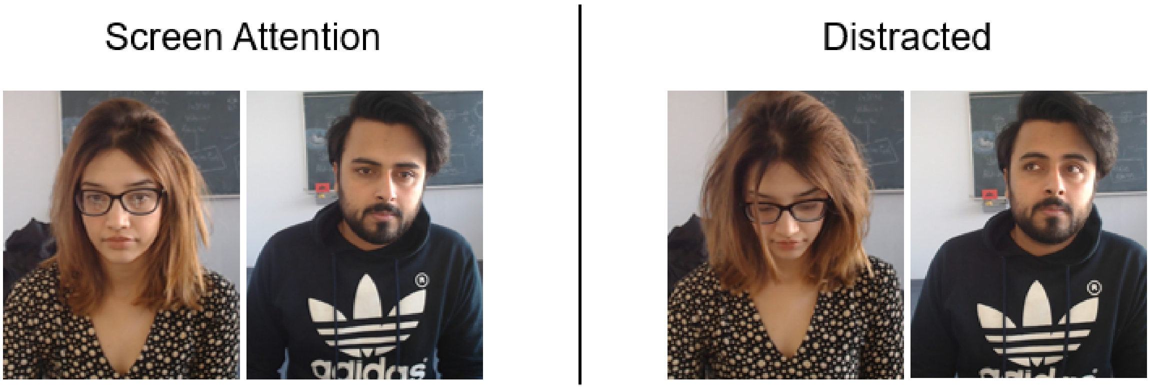


Figure 3.5: Sample images collected to differentiate between screen attention (on the left) and distraction (on the right). Images displayed here with special permission from the participants. The copyright remains with the dataset creators [Arora et al., 2024].

Gaze Mapping

The idea is to fine-tune the gaze estimation network for mapping gazes to either the screen region or elsewhere (indicating distraction). This process involves a transfer learning approach, where the weights learned by the gaze estimation model (discussed in Section 3.3.1) are reused. Specifically, the convolutional layers’ weights are copied and frozen to prevent further modifications. The prediction layer is adjusted for two-class classification using Softmax activation. The fine-tuning process employs an SGD optimizer with a learning rate of 0.01 and Categorical Cross-Entropy loss. The gaze mapping process described above is depicted in Figure 3.6.

The input images go through a face detection step using MediaPipe face detection [Bazarevsky et al., 2019]. The detected faces are cropped to the face region, scaled to 224 × 224, and processed in batches of 15. To augment the training data, brightness in input images is randomly adjusted within the range of ± 25%. The model undergoes 50 epochs of training, and the best model across all epochs is saved.

A leave-one-subject-out (LOSO) validation technique is utilized. In this approach, data from four participants are used for training, leaving one for validation. This process is repeated for all the participants, yielding five models. Attention recognition performance is

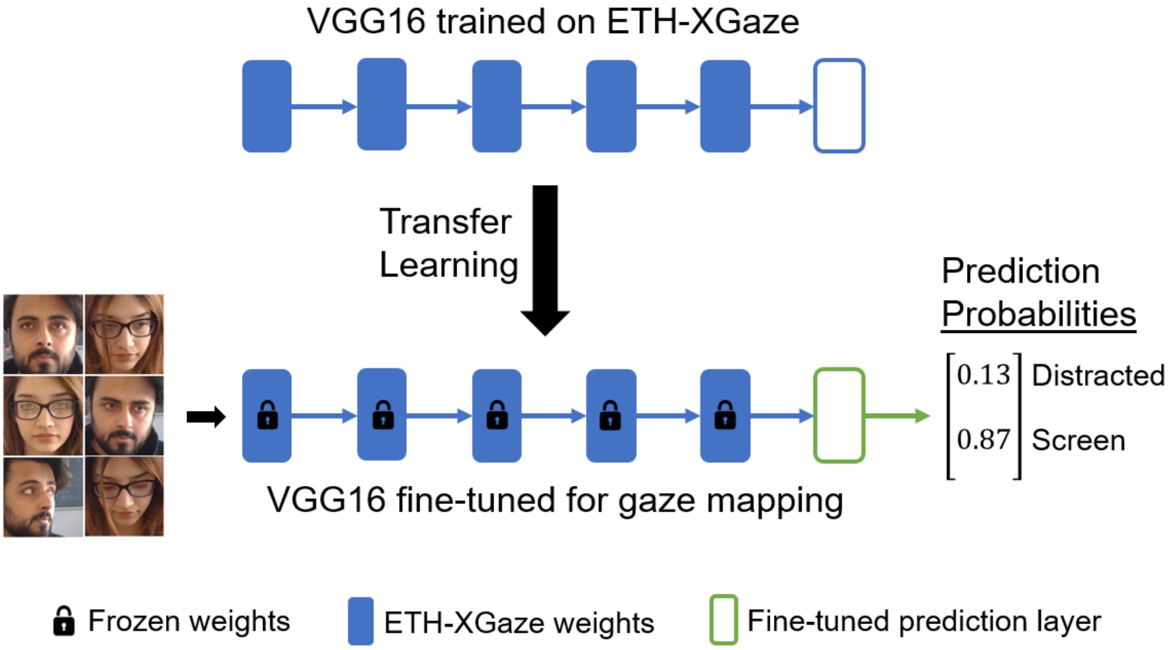


Figure 3.6: An illustration of the transfer learning process used to map gaze direction to attention classes (screen and distraction).

Model	Accuracy	F1-score
Participant 1	0.71	0.70
Participant 2	0.95	0.95
Participant 3	0.74	0.73
Participant 4	0.92	0.92
Participant 5	0.91	0.91
Average	0.846	0.842

Table 3.3: Results of LOSO Evaluation for Use Case 1 (Screen vs. Distracted). Participant $\langle i \rangle$ refers to the model obtained by leaving the i^{th} participant from training.

measured by accuracy and F1-score. As seen from Table 3.3, the models achieve an average accuracy of 84.6% and an F1-score of 84.2%. Interestingly, LOSO models for participants without glasses (Participants 2 and 4) outperform the average, with mean accuracy and F1-score reaching 93.5%. This is likely because, in certain cases, the participants' glasses had reflections from the screen that occluded the eye region.

3.3.3 Use Case 2 – Assembly Scenario

Use case 1 demonstrated the feasibility of mapping gaze directions using a small dataset. In this use case, the proposed approach is applied to an assembly scenario. The model developed here is further validated in Section 3.4 and subsequently utilized for participant behavior analysis and real-time robot behavior adaptation in Chapter 8.

Setup

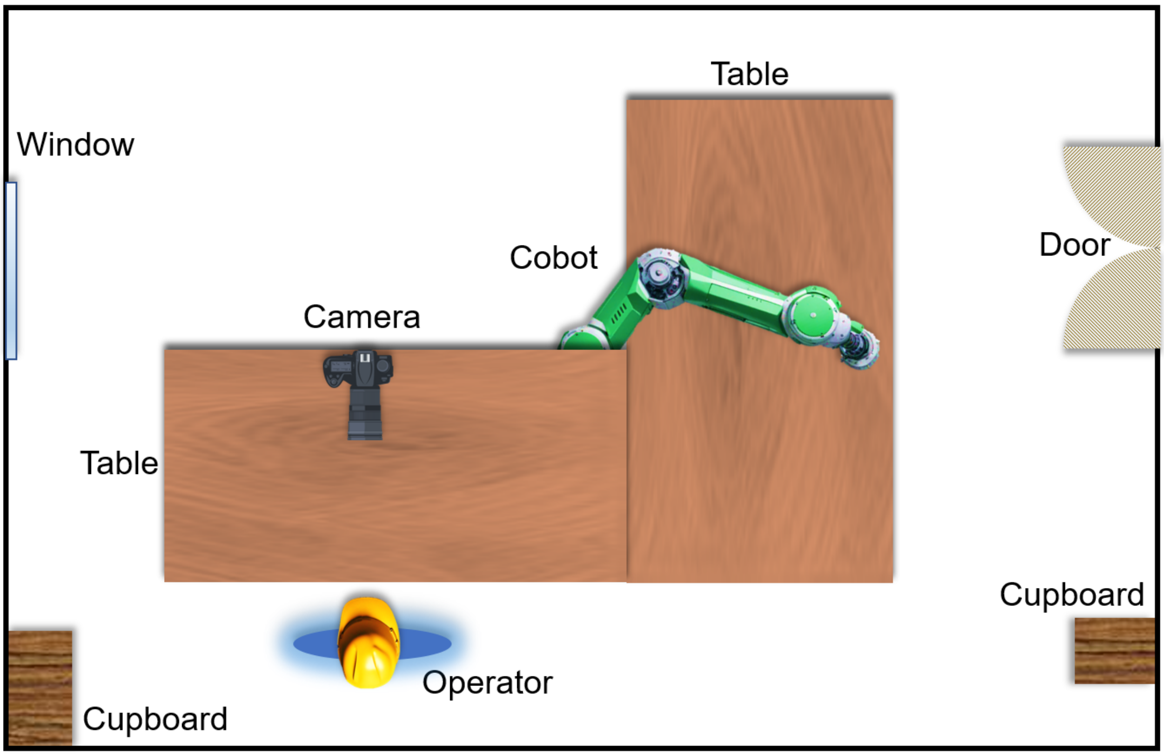


Figure 3.7: A visual representation of the setup layout highlighting the regions or objects of interest for attention and distraction recognition. This layout is reused with minor adjustments in the studies presented in Part 3 of this thesis.

The setup involves a human operator collaborating with a cobot to assemble a gearbox. However, in developing the attention recognition model for this setup, the primary focus is on identifying the key gaze areas, or areas of interest, rather than the specific assembly task. The setup's layout is depicted in Figure 3.7. The arrangement consists of tables configured in an L-shape, with the cobot positioned to the right of the operator. The operator carries out the various assembly steps on the table in front of them. It is important to highlight that the layout of the setup described in this section is nearly identical to the studies in Part 3 of this thesis. However, aspects of the setup like locations of windows and doors are critical to distraction detection but not other studies. So, these details have been removed from subsequent layout visualizations.

While piloting this setup, three main areas of interest were identified: the cobot, the assembly table, and the non-assembly areas. Non-assembly areas encompass spaces that are not directly related to the assembly task but may serve as potential sources of distraction. These non-assembly areas are distributed throughout the room, such as windows and cupboards.

Dataset

The dataset for the assembly scenario consists of images obtained from 8 adult volunteers (comprising 3 females and 5 males, aged 18 - 34 years). To capture these images, a Logitech

C920 Pro HD camera is positioned approximately 1.5 meters away from the operator, as illustrated in Figure 3.7. The data collection process follows a guided gazing protocol in which each participant is instructed to stand in front of the camera. They are then instructed to direct their gaze on one of three distinct areas: the cobot, the work table, or any other location in the room. Similar to Use Case 1, the participants are requested to vary their head orientation and gazing angles during the data acquisition. For each of the three specified conditions, 30 images are collected from each participant, resulting in a total of 720 labeled images with a resolution of 1920 × 1080 pixels. Figure 3.8 shows some example images from the collected dataset.

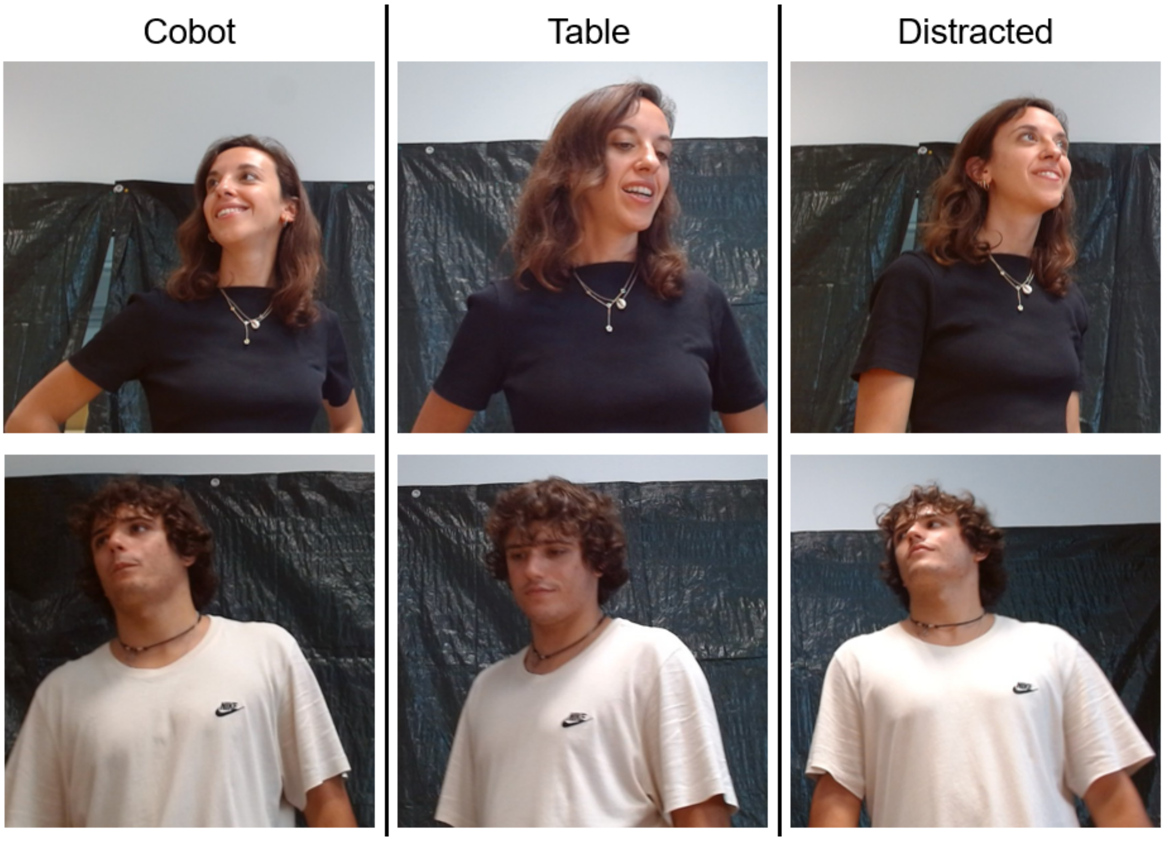


Figure 3.8: Some example images belonging to cobot, table, and distracted classes. The copyright remains with the dataset creators [Prajod et al., 2023a].

Gaze Mapping

Similar to Use Case 1 (Section 3.3.2), the approach in this scenario involves employing a transfer learning technique for fine-tuning the gaze estimation model. In this context, the prediction layer of the model is adapted to classify the input image into three defined classes: cobot, table, and distracted. The above dataset was used to map the gaze direction to these specified areas. The input image pre-processing, data augmentation techniques, and training parameters remain the same as those used in Use Case 1.

As with Use Case 1, models were trained using the LOSO validation method. This method yields a total of 8 models, each corresponding to one of the 8 participants. Performance measures such as class-wise recall, accuracy, and F1-score were computed to

Model	Recall			Accuracy	F1-score
	Cobot	Table	Distracted		
Participant 1	1.0	1.0	0.90	0.97	0.97
Participant 2	0.97	0.87	0.90	0.91	0.91
Participant 3	1.0	1.0	0.97	0.99	0.99
Participant 4	0.97	1.0	0.97	0.98	0.98
Participant 5	0.93	1.0	0.93	0.96	0.95
Participant 6	0.90	1.0	0.89	0.93	0.93
Participant 7	0.83	0.90	0.74	0.83	0.82
Participant 8	1.0	1.0	0.90	0.97	0.97
Average				0.943	0.94

Table 3.4: Results of LOSO Evaluation for Use Case 2 (Cobot vs. Table vs. Distracted). Participant $\langle i \rangle$ refers to the model obtained by leaving the ith participant from training.

provide a comprehensive assessment of the classification performance of these models. The results of LOSO validation are presented in Table 3.4. Notably, all models performed well in LOSO validation, achieving an average accuracy of 94.3% and an average F1-score of 94%. It's worth mentioning that the model corresponding to Participant 3 outperforms the rest, achieving an accuracy of 99% and an F1-score of 99%. On the other hand, the model corresponding to Participant 7 exhibits comparatively lower performance with an accuracy of 83% and an F1-score of 82%. Manual inspection of the images from Participant 7 revealed that certain images exhibited blurriness due to head movement, while others suffered from inaccuracies in face boundary detection, as illustrated in Figure 3.9.

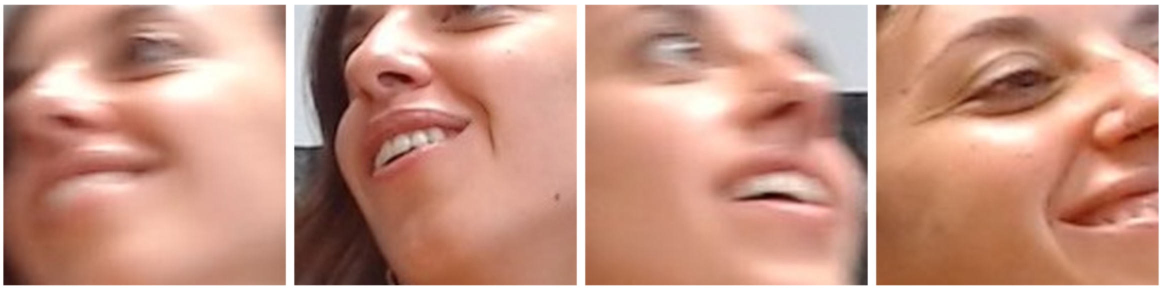


Figure 3.9: Illustrative images from Participant 7 showing blurriness and inaccurate face detection. The copyright remains with the dataset creators [Prajod et al., 2023a].

3.4 Validating Model in Industry-like HRC

As demonstrated in Sections 3.3.2 and 3.3.3, the utilization of transfer learning for mapping gaze direction yielded models that perform well in attention and distraction recognition. The outcomes of the LOSO evaluation highlight the robustness of these models when confronted with images of previously unseen participants. Nevertheless, it is important to note that all the images used in these evaluations originate from the same dataset, which was collected under a guided gaze protocol. In this case, the attention and distraction states captured are intentionally simulated rather than naturally elicited in the participants.

To gain deeper insights into how well these models can perform in detecting genuine instances of attention and distraction, a new dataset was collected. This dataset is derived from videos of participants engaging in a collaborative assembly task with a cobot. Unlike the previous datasets, these videos were recorded during an extended study, making them more likely to capture unguided, natural behaviors that manifest during genuine attention and distraction. This new dataset is instrumental in validating the attention recognition model described in Section 3.3.3, ensuring that its performance extends to real-world scenarios characterized by unscripted, natural behavior.

3.4.1 Data Acquisition

The validation dataset was collected within a controlled laboratory environment designed to emulate an Industry 5.0 assembly process. To achieve this, the same setup described in Section 3.3.3 was employed. Video recordings were obtained from 8 participants, consisting of 3 males and 5 females with ages ranging from 18 to 30 years. Each participant was assigned the role of an operator and engaged in the task for 3.5 hours each day, continuously for 5 consecutive days, effectively simulating the workweek experience of a cobot worker.

Given the extended duration of the study, spanning multiple days, it was anticipated to naturally elicit instances of genuine attention and distraction. The assembly task itself encompassed four primary phases: Gathering Components, Assembling, Waiting for Cobot, and Collaborative Joining of Sub-assemblies, all of which together constitute a single assembly cycle, as depicted in Figure 3.10. The primary objective of this data collection was to create a dataset that represented naturalistic gaze behaviors. Hence, participants were not provided with guided or scripted gaze instructions. Instead, the data acquisition process was designed to encourage participants to exhibit their own, unguided gaze behaviors.

The videos were recorded using the frontal camera (see Figure 3.7), at a resolution of 1280 × 720 and 25 frames per second. During the first workday, three videos of approximately 10 minutes each were recorded. These videos provided a representative sampling of the entire workday, comprising segments from the beginning, middle, and end of the day. Likewise, three additional videos were recorded during the last workday of the experiment. This approach led to the acquisition of one hour of video data for each participant, resulting in a total of 8 hours of recorded content. The dataset acquisition followed the guidelines of the Declaration of Helsinki and was approved by the Ethics Committee of I.R.C.C.S. Eugenio Medea (protocol code N. 19/20—CE of 20 April 2020). The dataset is utilized to validate the gaze-based attention recognition model in a context involving unscripted, natural behaviors.

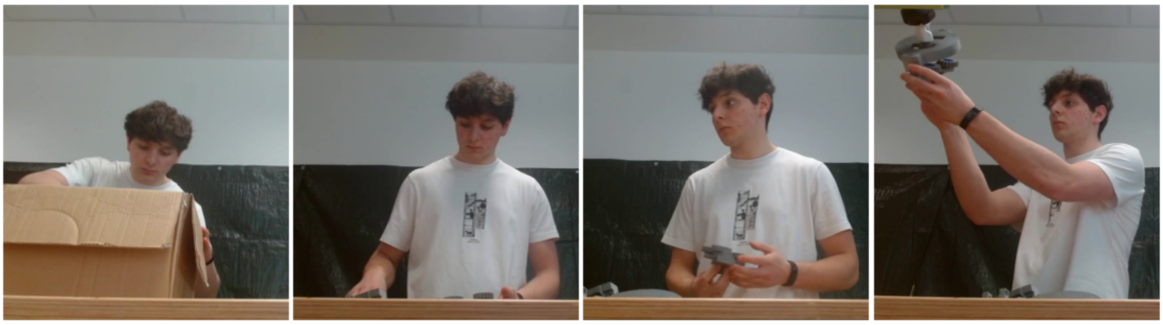


Figure 3.10: A series of images illustrating the collaborative assembly task. From left to right, the operator performs various phases - gathering components, assembling, waiting for the cobot, and finally, collaboratively joining sub-assemblies with the cobot. The copyright remains with the authors [Prajod et al., 2023a].

3.4.2 Validation Dataset

The videos were annotated to align with the various phases of the assembly task that participants engaged in. This approach enables the identification of segments where the participant's attention is directed towards the table, the cobot, or when they are distracted. For instance, during the Assembling phase, participants typically focus their gaze downward at the table since this is where their assembly work takes place. However, during the Gathering Components phase, the placement of the component box influences the participant's visibility. Some participants opted to place the box on the floor, which required them to bend down to retrieve components, rendering them out of the camera's view. So, only the Assembling phase could reliably provide samples of attention to the table.

Although the assembly task was shared equally between the cobot and the human operator, there was a considerable difference in their respective production speed. This discrepancy led to extended waiting periods for the operator. During these waiting segments, the operator may either direct their attention towards the cobot or look in random directions, such as checking their watch or looking out of a window. Consequently, these Waiting segments serve as samples for identifying attention to the cobot or moments of distraction. In addition to the Waiting segments, the segments from Collaborative Joining phase were also considered for attention to cobot. However, the participants often lifted their hands in front of their faces while performing the joining activity, leading to their faces being occluded in the camera view.

Each labeled video segment corresponds to a short duration, typically lasting only for some seconds. Notably, there is minimal variation in gaze behavior within a given segment. Consequently, to avoid repetition, three representative frames were extracted from each segment: first, middle, and last frames. Following this extraction process, the selected images underwent a pre-processing step. In this step, the presence of a human face was detected within the image, and the images were cropped to the facial region. It's worth noting that any images where face detection failed were discarded from the dataset. Additionally, images that exhibited blurriness, often due to sudden quick movements during the recording, or where the participant's eyes were obscured by objects, were also omitted. In total, approximately 20% of the initially selected images were excluded based on these

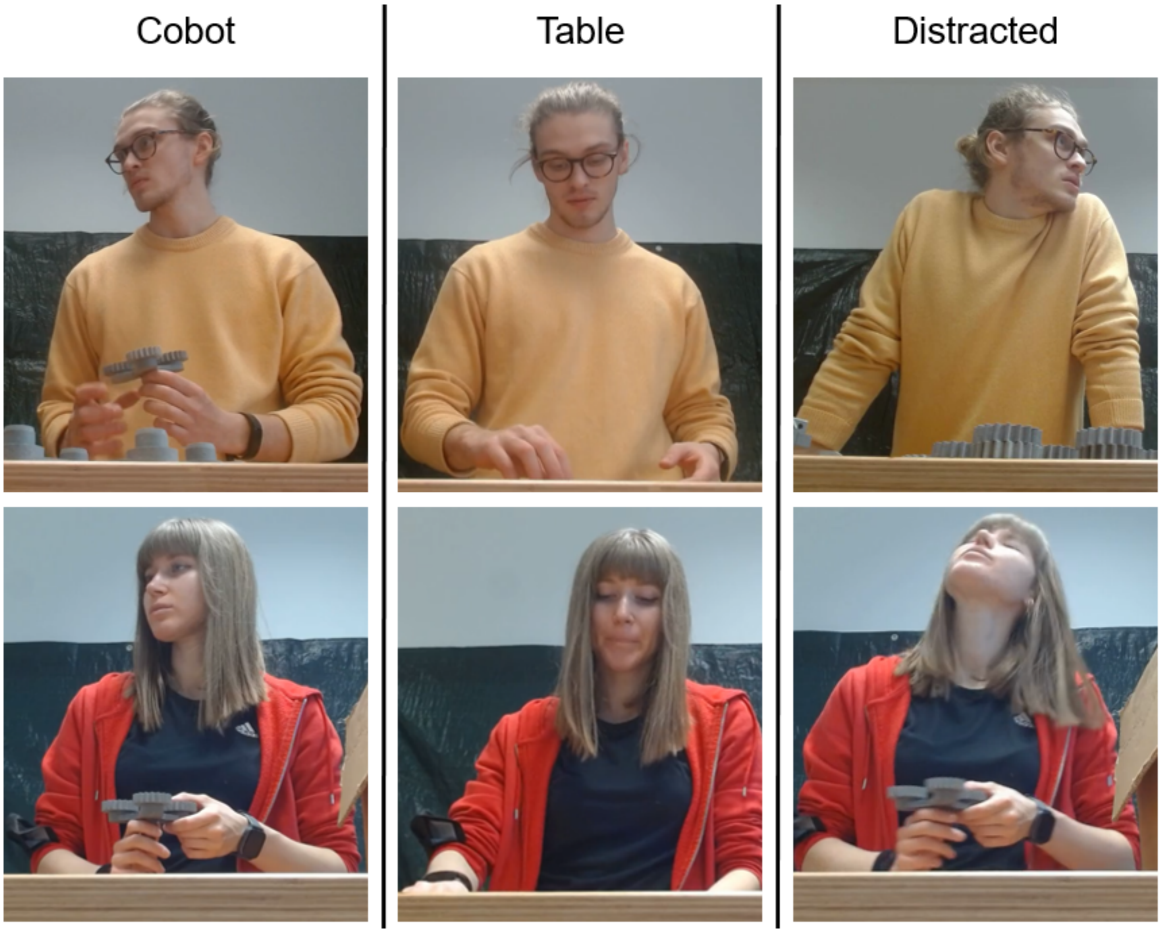


Figure 3.11: Some sample images from the validation dataset. From left to right - Attention to the cobot (while waiting), Attention to the table (while assembling), and Distracted (while waiting). The copyright remains with the dataset creators [Prajod et al., 2023a].

criteria. Consequently, the resulting validation dataset consists of 833 images indicating attention directed towards the cobot, 940 images depicting attention focused on the table, and 962 images representing moments of distraction. Figure 3.11 provides examples of images from this dataset, each labeled to indicate the participant's state of attention.

3.4.3 Results and Discussion

The 8 LOSO models previously discussed in Section 3.3.3 were evaluated in an industry-like collaborative assembly setting, using the validation dataset described in Section 3.4.2. The results of this cross-dataset evaluation are presented in Table 3.5. All the models have similar performance and achieve an accuracy and F1-score of 81 - 82%. Although the recalls are lower than Use case 2, the performance of the models indicate the applicability of these models to industrial settings.

In all models, the drop in performance mainly stemmed from the Distracted class, followed by the Attention to Cobot class. To gain insights into this reduction in recall, an investigation was conducted on the confusion matrix of predictions generated by the models. Figure 3.12 shows the confusion matrix corresponding to the predictions from the Partic-

Model	Recall			Accuracy	F1-score
	Cobot	Table	Distracted		
Participant 1	0.85	0.98	0.61	0.81	0.81
Participant 2	0.87	0.95	0.66	0.82	0.82
Participant 3	0.87	0.95	0.65	0.82	0.82
Participant 4	0.83	0.95	0.67	0.81	0.82
Participant 5	0.89	0.98	0.61	0.82	0.82
Participant 6	0.86	0.96	0.62	0.81	0.81
Participant 7	0.87	0.96	0.63	0.82	0.82
Participant 8	0.83	0.94	0.67	0.82	0.82
Average				0.816	0.818

Table 3.5: Performance of the eight LOSO models from Use Case 2 on the validation dataset.

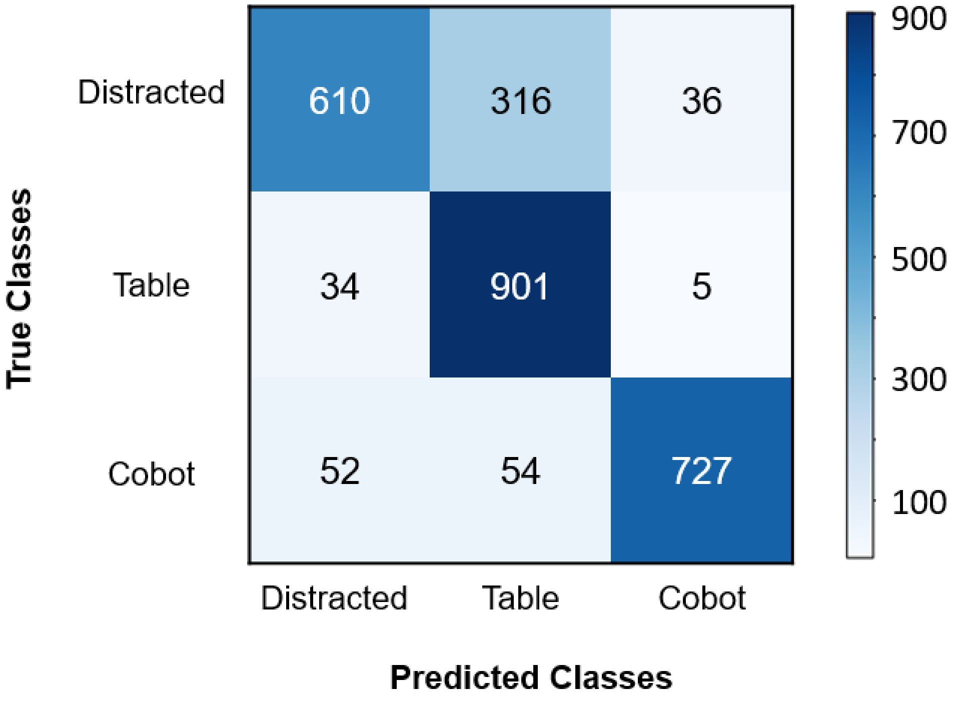


Figure 3.12: Confusion matrix for Participant 7 model predictions on the validation dataset

ipant 7 model on the validation dataset. It's worth noting that all models exhibit similar confusion matrices, and interestingly, most of the misclassified Distracted images are predicted as Attention to Table.

The observed trend was further explored through a manual inspection of misclassified images. Notably, many instances revealed participants being distracted by items located on the table. For instance, during periods of waiting, participants often directed their gaze towards a sub-assembly on the table. Some illustrative examples of Distracted images, where participants are looking in the direction of the table even when they are not actively assembling parts, are shown in Figure 3.13. Observations from the manual inspection indicate that the classification performance could potentially be enhanced by incorporating additional data, such as the proximity of the hand to the table and the body pose of the operator.

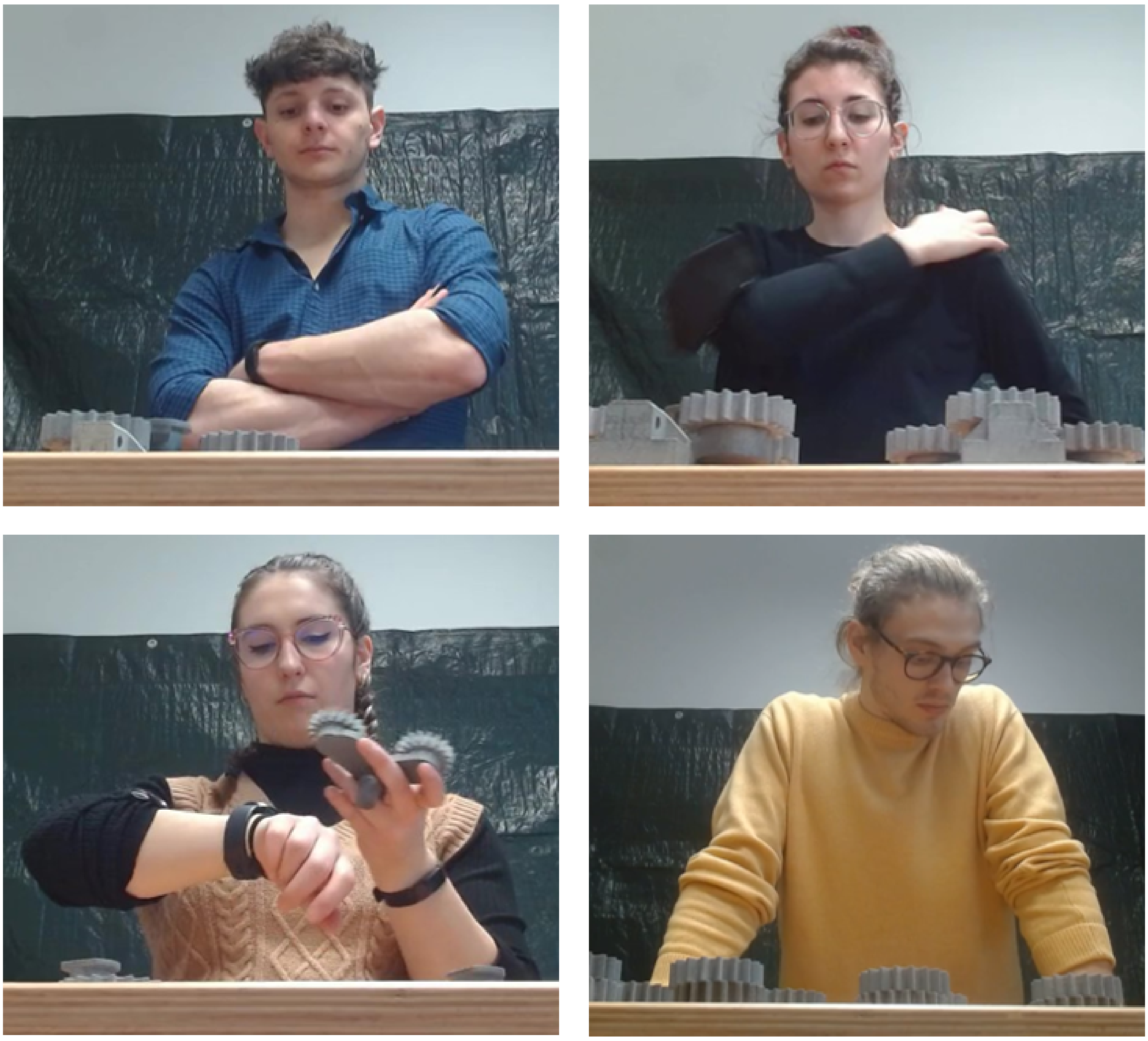


Figure 3.13: Few examples of images belonging to the Distracted class that were misclassified as Attention to Table. Reused with permission, the copyright remains with the authors [Prajod et al., 2023a].

3.5 Reflections and Remarks

This chapter delves into the domain of attention and distraction recognition in the context of human-robot collaboration, with a specific focus on the Industry 5.0 context. It addresses two challenges faced in this field: the limitations imposed by small dataset sizes and the necessity for model validation in natural, unscripted settings. The models developed and discussed in Sections 3.3.2 and 3.3.3 demonstrate that, even with relatively limited datasets, it is feasible to map gaze effectively for a given industrial setup. This finding is especially significant in practical scenarios where acquiring large datasets tailored to the specific setup can be a challenging task. Furthermore, the research presented in Section 3.4 highlights the pressing need to shift from guided, scripted data collection methods to unscripted, natural scenarios for model validation. This validation is necessary for evaluating the applicability and effectiveness of attention and distraction recognition models within industrial settings.

Chapter 4

Pain Detection

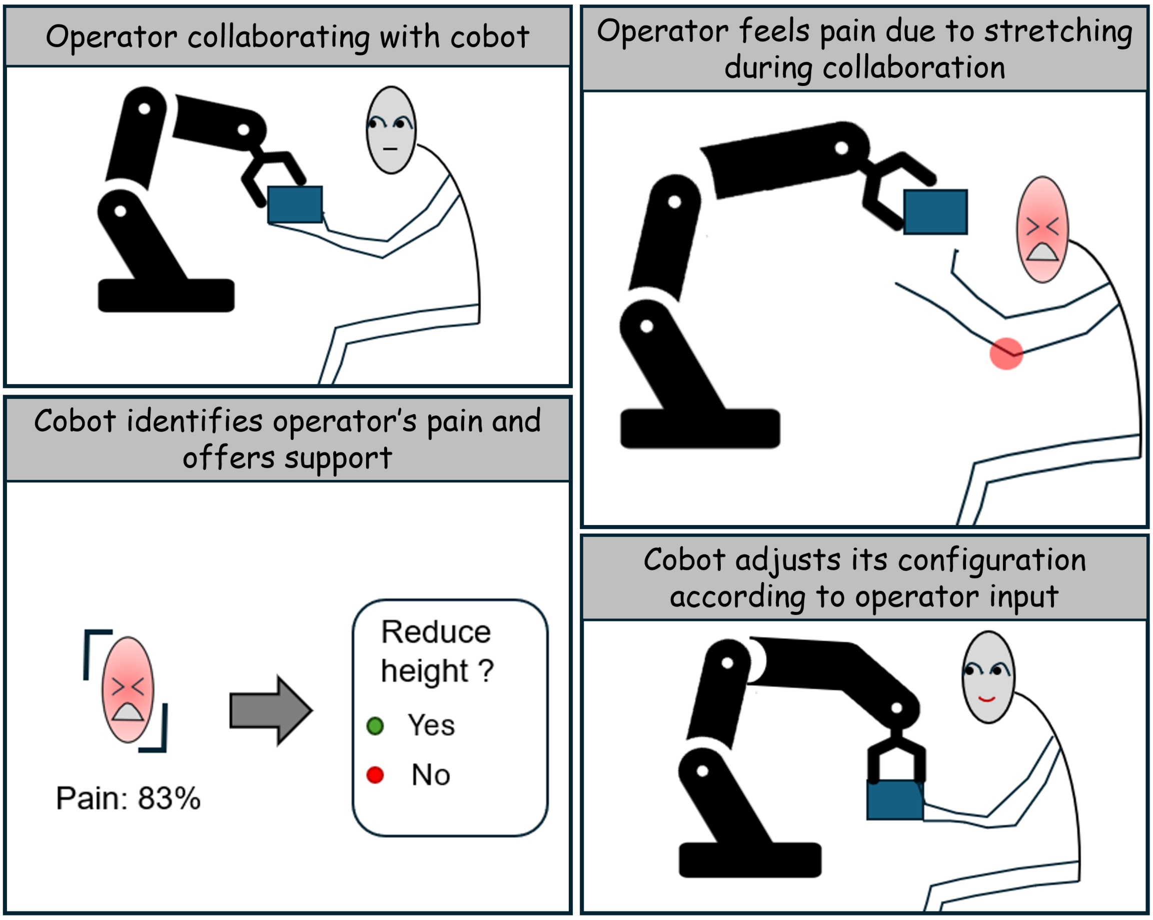

Operator collaborating with cobot
Operator feels pain due to stretching during collaboration
Cobot identifies operator's pain and offers support
Pain: 83%
Reduce height ?
● Yes
● No
Cobot adjusts its configuration according to operator input

Figure 4.1: A comic strip illustration of how pain detection can improve the collaboration experience in an industrial scenario. In this situation, an operator experiences pain due to a certain stretching movement. The cobot detects the operator's pain expression and offers to modify its configuration to avoid stretching. The cobot proceeds to change its configuration according to the operator's preference.

4.1 Overview

In the context of Industry 5.0, ensuring the well-being of workers is essential for a safe and efficient workplace. Pain detection is integral for identifying and addressing discomfort or injuries resulting from workplace conditions. Imagine the scenario depicted in Figure 4.1, where an operator experiences pain while stretching hands for collaboration. The cobot detects the pain expression and proactively offers to modify its configuration. This not only addresses the operator's discomfort but also fosters a more effective human-robot collaboration.

This chapter proposes a pain detection model developed through a transfer learning approach, leveraging feature representations initially learned for emotion recognition. Additionally, it addresses the often-overlooked aspect of quantitatively assessing the learned feature representations of the model. An approach based on eXplainable Artificial Intelligence (XAI) is introduced for this purpose. The approach is further employed to compare the feature representations of models trained for two different types of pain, facilitating the assessment of prominent features in specific datasets. This analysis helps determine whether a dataset can effectively be used to train robust and generic pain detection models. The content of this chapter builds upon and extends the research presented in:

- P. Prajod, D. Schiller, T. Huber, and E. André. Do deep neural networks forget facial action units?—Exploring the effects of transfer learning in health related facial expression recognition. *AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action*, 1013:217, 2022b

  [*I developed the machine learning models and performed the analyses. I also contributed significantly to the development of the analysis framework. Furthermore, I contributed significantly to formulating research questions and deriving insights.*]

- P. Prajod, T. Huber, and E. André. Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions. In *International Conference on Multimedia Modeling*, pages 311–322. Springer, 2022a

  [*I developed the machine learning models and performed the analyses. I also contributed significantly to formulating research questions and deriving insights.*]

4.2 Background Literature and Previous Works

This section provides a background for pain detection (Section 4.2.1) and an overview of two approaches - XAI visualizations (Section 4.2.2) and Cross-Dataset evaluations (Section 4.2.3) - to assess the generalizability of pain models.

4.2.1 Towards Automatic Pain Detection

Pain Detection in Industry 5.0

Detecting pain early enables proactive interventions in industrial settings, improving both job satisfaction [Yamada et al., 2016; Hoogendoorn et al., 2000; Baek et al., 2018] and productivity [Witt et al., 2016; Fan and Straube, 2016; Harman and Ruyak, 2005]. In the context of human-robot collaborations, the ability to detect pain becomes crucial for cultivating positive relationships between human operators and cobots. For instance, social support is a crucial psychosocial factor that influences work-related pain [Baek et al., 2018; Hoogendoorn et al., 2000], with potential implications for the development of chronic pain Matthias et al. [2022]. Currently, cobots cannot detect pain in operators, consequently limiting their ability to offer any form of social support.

Pain Datasets

As mentioned in Chapter 1, collecting pain datasets is challenging due to ethical concerns and patient availability. This section identifies existing pain datasets that recorded facial expressions of pain, which is the focus of this chapter. The identified datasets are presented in Table 4.1. Datasets such as Hi4D-ADSIP [Matuszewski et al., 2011] and Delaware [Mende-Siedlecki et al., 2020] pain datasets were not considered as they collected posed pain expressions.

Among the existing facial pain datasets, some [Brahnam et al., 2006; Heiderich et al., 2015; Egede et al., 2019; Yan et al., 2020; Brahnam et al., 2023] are collected from infants to develop neonatal pain assessment systems. They typically utilize medical procedures or pain-inducing punctures (e.g., pin picks, pinching) to elicit pain. Since self-reports are not feasible, pain is often labeled depending on the activity (e.g., before vs. during the procedure). Moreover, the datasets typically include labels for crying induced by non-pain stimuli (e.g., hunger, fear).

The other datasets listed in Table 4.1 are collected from adults undergoing painful stimuli. Notably, UNBC-McMaster [Lucey et al., 2011] and EmoPain [Aung et al., 2015] datasets are collected from participants with pre-existing conditions. Hence, their pain experience (e.g., shoulder pain, lower back pain) can be considered as manifestations of chronic pain. The other adult pain datasets such as BioVid [Walter et al., 2013] and MInt-PAIN [Haque et al., 2018] are collected from participants undergoing acute pain stimuli (e.g., heat, electrical currents). The UNBC-McMaster, BP4D-Spontaneous [Zhang et al., 2014], and PEMF [Fernandes-Magalhaes et al., 2023] datasets collected only image/video data, whereas others such as BioVid and X-ITE [Gruss et al., 2019] collected physiological signals (e.g., electrocardiogram, electrodermal activity, etc.) as well.

Pain Annotations and Predictions

The field of affective computing has proposed a variety of models for detecting, recognizing, and estimating pain from facial expressions. Inspired by the terminologies from facial Action Unit (AU) models, this thesis distinguishes between detection, recognition, and estimation. Pain detection involves binary classification of whether a facial expression exhibits pain or not. Whereas, pain recognition entails multi-class classification of pain levels, such as differentiating between no pain, mild pain, and severe pain. Pain estimation, on the other hand, predicts pain intensity on an ordinal scale (e.g., ranging from 0 to 15). However, comparing pain recognition and estimation models poses a significant challenge due to the absence of standardized pain scales and class labels. For instance,

Paper	Stimuli	Pain Scale
Brahnam et al. [2006] (iCOPE)	Pin prick and puncture	2 levels - activity-based labeling
Lucey et al. [2011] (UNBC-McMaster)	Shoulder pain	16 levels - AU-based pain intensity
Walter et al. [2013] (BioVid)	Heat pain	5 levels - Individual pain thresholds
Zhang et al. [2014] (BP4D-Spontaneous)	Cold pain	6 levels - Self reported
Aung et al. [2015] (EmoPain)	Lower-back pain	Continuous 0 to 1 observer ratings
Heiderich et al. [2015] (UNIFESP)	Medical procedure	2 levels - activity-based labeling
Velana et al. [2017] (Sense Emotion)	Heat pain	4 levels - Individual pain thresholds
Haque et al. [2018] (MIntPAIN)	Electrical pain	5 levels - Individual pain thresholds
Gruss et al. [2019] (X-ITE)	Heat and electrical pain	4 levels for each type - Individual pain thresholds
Egede et al. [2019] (APN-db)	Medical procedure, vaccination	12 levels - behavior-based pain intensity
Yan et al. [2020] (FENP)	Natural settings	3 levels - observer labeling
F-Magalhaes et al. [2023] (PEMF)	Pressure pain	9 levels - observer ratings
Brahnam et al. [2023] (iCOPEvid)	Pin prick and puncture	2 levels - activity-based labeling

Table 4.1: An overview of available pain datasets that collected facial images/videos.

the UNBC-McMaster shoulder pain dataset [Lucey et al., 2011] utilizes the PSPI (Prkachin and Solomon Pain Intensity) scale to assess pain in image sequences. This scale assigns a numerical pain level value between 0 and 15, which is computed based on facial AU annotations. In contrast, the BioVid heat pain dataset [Walter et al., 2013] relies on participant-specific pain tolerance levels to label pain intensity (0 - 4) in videos. Consequently, this chapter primarily focuses on pain detection. Nevertheless, the research gaps identified below extend beyond pain detection tasks and are equally relevant to pain recognition and estimation tasks.

XAI Methods

Recent advancements in pain facial expression prediction have placed a strong emphasis on explaining the model's decision-making process. A widely employed approach in health-related domains is visual explanation [van der Velden et al., 2022]. This approach involves highlighting specific pixels or regions in the input image that contributed most significantly to the prediction. During the development of pain models, generating and evaluating explanations are essential to ensure that the models extract meaningful facial features that occur in pain expressions. The pain models reviewed in the next section (Section 4.2.2) predominantly utilize four XAI methods to generate heatmaps (also called saliency maps): GradCAM (Gradient-weighted Class Activation Mapping) [Selvaraju et al., 2017], LRP (Layer-wise Relevance Propagation) [Bach et al., 2015], LIME (Local Interpretable Model-agnostic Explanations) [Ribeiro et al., 2016], and SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017].

The Grad-CAM method computes the gradient of the target class score with respect to the feature maps of a convolutional layer. It uses a weighted combination of gradients from the feature maps to produce a coarse heatmap that highlights the salient regions in the image for the target class.

The LRP technique propagates the output prediction backward through the network, distributing the prediction score to each neuron based on their contribution. It provides a pixel-wise decomposition of the model's decision, producing a fine-grained heatmap highlighting input features that are most important for a particular prediction.

LIME is a model-agnostic method that explains individual sample predictions of a classifier by approximating the model's behavior using a simpler model. It generates interpretable explanations by perturbing the input and observing how the model's output changes. Like Grad-CAM, this method also produces coarse visualizations.

The SHAP algorithm calculates the Shapley value for each feature, which represents the average contribution of a feature to the prediction across all possible combinations. The sum of the calculated SHAP values for all features equals the model's prediction. This technique is model-agnostic and produces fine-grained visualizations.

4.2.2 XAI-based Investigation of Pain Models

Existing pain detection models that utilize XAI methods can be broadly categorized based on their intended purpose as: neonatal pain detection and pain detection in adults. Neonatal pain detection models typically utilize datasets such as iCOPE and UNIFESP datasets for training and evaluation. On the other hand, pain recognition models for adults often employ datasets like UNBC shoulder pain and BioVid heat pain for their model development. This chapter focuses exclusively on pain datasets involving adult participants because the aim is to develop a robust pain detection model for industrial settings. The following works were identified through Scopus[1] literature search for their application of XAI methods in their pain prediction models.

[1] `https://www.scopus.com/`, Query: TITLE-ABS-KEY ((face OR facial) AND pain AND (deep*learning OR ml OR machine*learning OR network) AND (detect* OR recogni* OR predict* OR estimat* OR classif*) AND (explaina* OR interpretab* OR salien* OR relevan* OR xai))

One of the first works to employ XAI techniques to explain pain prediction was presented by Xu et al. [2019]. They employed multitask learning to simultaneously recognize self-reported Visual Analog Scale pain scores and AUs from the UNBC shoulder pain dataset. They utilized the SHAP method to generate saliency maps, highlighting pixels in varying intensities based on their impact on the model's prediction. Their results indicated that the model captured facial regions associated with pain expression, particularly the eyebrows, mouth, and nasolabial folds. These regions also influenced the prediction of specific pain AUs, such as the eye area in AU7 and AU43, and the mouth in AU25.

Various XAI methods are available for generating explanations. But, which method is suitable for a particular explanation objective? To address this question, Weitz et al. [2019] conducted a comparative study of two explanation methods - LRP and LIME - and assessed their effectiveness in explaining a model's decision and identifying relevant facial regions. They fine-tuned a VGG-Face [Parkhi et al., 2015] network on the BioVid heat pain dataset to discern between Happy, Disgust, and Pain expressions. They found that LRP produced more granular saliency maps, while the visualizations generated through LIME made it easier to identify relevant facial regions. While examining the visualizations for the pain class, they did not find a consistent facial pattern across images. However, they observed that these visualizations highlighted background features or irrelevant features such as hair and neck, which they believe partly accounts for the model's poor performance.

Rieger et al. [2020] underscored the significance of considering explanations alongside classification performance in evaluating pain detection models. They trained a ResNet [He et al., 2016] network to detect pain-related AUs using the Actor Study [Seuss et al., 2019] and Extended Cohn-Kanade [Lucey et al., 2010] datasets, and evaluated it on the UNBC shoulder pain dataset. By generating LRP saliency maps, they identified the pixels that contributed most significantly to the prediction of each AU. They then used facial landmarks to determine the bounding boxes for each AU. They observed that, in the majority of cases, the highlighted pixels were located outside the corresponding bounding boxes.

Semwal and Londhe [2021b] trained a model to recognize pain levels (0 - 5) using a combination of RGB images and texture images to obtain features from both modalities. The model was trained on a dataset collected by the authors from 10 participants in a hospital setting. They generated CAM heatmaps [Zhou et al., 2016] to visualize the regions that the model focused on for different pain levels. Their findings revealed that for all pain levels, the model's attention was primarily concentrated on the face, particularly the forehead, eye, nose, and mouth regions. Additionally, the model placed less emphasis on the background region.

Yuan et al. [2022] developed a transformer-based pain recognition model using the UNBC shoulder pain dataset. They introduced the concept of pre-training the model with masked faces from the same dataset. To visualize the model's focus, they employed the Grad-CAM method to generate heatmaps. They observed that without mask pre-training, the model primarily focused on the eye and eyebrow regions. However, they noted that while eye features are effective in discerning high-pain expressions, they may not be sufficient for distinguishing between lower pain levels. Notably, applying an upper-region mask during pre-training led to a shift in the model's attention towards nasolabial folds and mouth corners. This result suggests that mask pre-training enabled the model to incorporate a broader range of pain-related facial features.

In the context of neonatal pain detection, Carlini et al. [2021] fine-tuned a VGG-Face network utilizing images from both the iCOPE and UNIFESP datasets. Using the gradient [Sundararajan et al., 2017] technique, they generated pixel-wise saliency maps for the test images. Their findings revealed consistent highlighting of key facial features such as the forehead, upper nose contour, and mouth with tongue protrusion. However, they noted that some images from the UNIFESP dataset highlighted pixels extending beyond the face (e.g., blanket) – an observation they attributed to the dataset's lower recording quality compared to iCOPE.

Coutrin et al. [2022] fine-tuned four existing Convolutional Neural Networks (CNNs) for pain detection using the iCOPE and UNIFESP neonatal pain datasets. Additionally, they trained the N-CNN network from scratch without leveraging transfer learning. Heatmaps generated using Grad-CAM revealed that only one of the models (VGG-Face) consistently highlighted facial regions, such as the nose, nasolabial region, and forehead, which are known markers of pain expression. Notably, the other fine-tuned models either focused on the entire face or detected irrelevant non-facial features (e.g., hair ornaments). The authors observed that the N-CNN also highlighted facial regions, but there was no consistent pattern across heatmaps. They attributed the N-CNN model's inconsistent heatmaps to the limited amount of training data and lack of transfer learning.

One of the widely adopted network architectures for neonatal pain assessment (N-CNN) was proposed by Zamzmi et al. [2019]. In their study, Ferreira et al. [2023] investigated hyperparameters such as image size, optimizer, and epochs to identify optimal training parameters for the neonatal pain detection task. Similar to Carlini et al., they also utilized both iCOPE and UNIFESP datasets to train the model. To generate explainable heatmaps, they utilized both Grad-CAM and IG methods. Their model focused on relevant facial regions and exhibited lesser focus on background features like clothing. Furthermore, their model exhibited a more uniform distribution of highlighted regions across the face, unlike the non-tuned model, which heavily emphasized the eyes and mouth. Notably, they observed that pain predictions with a confidence of 40-60% highlighted features unrelated to the face. In contrast, predictions at the extreme ends (less than 20% or greater than 80%) highlighted more relevant facial features.

Building upon their previous research on neonatal pain detection models [Carlini et al., 2021; Coutrin et al., 2022], Carlini et al. [2024] examined the Grad-CAM and IG heatmaps generated by VGG-Face and N-CNN models. Similar to their earlier studies, they employed the iCOPE and UNIFESP datasets for training and evaluation. The N-CNN model's Grad-CAM and IG heatmaps consistently highlighted the mouth area in pain images, while the VGG-Face model's heatmaps focused on regions like the forehead, eyes, nose, and mouth. Additionally, they analyzed the visual attention of human participants by tracking their gaze patterns while viewing the images. Their correlation analysis revealed highest overlap between VGG-Face's Grad-CAM heatmaps and human assessments.

A closer examination of the literature revealed two nuanced objectives for generating explanations in the context of pain prediction models. The first objective aimed to ensure that the model focused on the facial region and not on irrelevant background elements. In this case, explanations served as a tool to assess whether the model had mistakenly identified non-facial features like hair ornaments, neck, or other background elements as indicators of pain. The second objective focused on identifying specific regions of the face that played a crucial role in pain prediction. Some studies extended this objective to correlate the identified regions with pain-related AUs and facial patterns. Coarse-grained heatmaps, such as CAM and LIME, are well-suited for highlighting facial regions. However, pixel-level fine-grained heatmaps, like LRP and IG, are more suitable for identifying subtle facial features like AUs.

	Paper	XAI Method	Model Interpretation
NEONATAL	Carlini et al. [2021]	IG	Manual
	Coutrin et al. [2022]	Grad-CAM	Manual
	Ferreira et al. [2023]	Grad-CAM, IG	Manual
	Carlini et al. [2024]	Grad-CAM, IG	Manual
ADULT	Xu et al. [2019]	SHAP	Manual
	Weitz et al. [2019]	LRP, LIME	Manual
	Rieger et al. [2020]	LRP	Manual, Counting
	Semwal and Londhe [2021b]	CAM	Manual
	Yuan et al. [2022]	Grad-CAM	Manual
	Prajod et al. [2022b]*	**LRP + TCAV**	Manual, **Statistical**
	Prajod et al. [2022a]*	**LRP + TCAV**	Manual, **Statistical**

*Table 4.2: An overview of the existing works on predicting pain from facial expressions that employ XAI techniques, along with the model's heatmap interpretation method. The entries marked with * are expanded in the subsequent sections of this chapter.*

4.2.3 Assessing Generalizability of Pain Models

The findings from XAI techniques can serve as indicators of a model's generalization capabilities [Doshi-Velez and Kim, 2018]. For instance, a model that relies on background or non-facial information for its predictions will likely struggle when applied in a new environment with a different background. However, as evident from Table 4.2, insights obtained from XAI heatmaps are predominantly based on manual inspection and often lack quantitative analysis.

A commonly used approach for assessing the generalizability of a pain model is to reserve data from some unseen participants for testing, such as leave-one-subject-out and participant hold-out methods [Hassan et al., 2019; Gkikas and Tsiknakis, 2023a]. However, the unseen participants in the test set are subject to the same pain stimuli and are recorded under identical setups. That is, within-dataset evaluations do not evaluate how well the model performs in detecting pain in users subject to different pain stimuli or recorded under varying conditions. As demonstrated by Dai et al. [2019], such models may learn dataset-specific features and fail to generalize effectively to other pain datasets.

	Paper	Objective
NEONATAL	Carlini et al. [2021]	Visualize facial features, Face vs. background
	Coutrin et al. [2022]	Compare face regions focused by different models
	Ferreira et al. [2023]	Shift in focused regions due to hyperparameter tuning, Face vs. background
	Carlini et al. [2024]	Compare CNN-generated explanations with human visual attention
ADULT	Xu et al. [2019]	Visualize face features
	Weitz et al. [2019]	Comparing XAI methods on interpretability
	Rieger et al. [2020]	Visualize AU regions, Highlighted pixels within AU bounding box
	Semwal and Londhe [2021b]	Visualize facial features, Face vs. background
	Yuan et al. [2022]	Shift in focused regions due to masking
	Prajod et al. [2022b]*	**Shift in relevance of AUs due to transfer learning**
	Prajod et al. [2022a]*	**Compare facial features learned by two distinct models**

*Table 4.3: An overview of the existing facial pain detection models that generate visual explanations and their objectives for employing XAI techniques. The entries marked with * are expanded in the subsequent sections of this chapter.*

To address this limitation, cross-dataset evaluations have been employed to expand the assessment of a model's generalization capabilities. This approach involves training a model on a dataset A and evaluating its performance on a different dataset B. The model is deemed to generalize well if it yields consistent performance on both datasets. The following works were identified through Scopus[2] literature search as works that evaluated their pain models using cross-dataset validations.

Recognizing the limited cross-dataset evaluations in pain detection research, Othman et al. [2019] conducted cross-dataset evaluations of their pain models. They trained two pain models, an RFC, and a deep learning model, each on both the BioVid heat pain dataset and the X-ITE thermal and electrical pain dataset [Gruss et al., 2019]. Subsequently, they evaluated the BioVid models on the X-ITE dataset and vice versa. Both the BioVid and X-ITE models achieved comparable performances in both within-dataset and cross-dataset evaluations. This led the authors to conclude that both models demonstrated strong generalization capabilities.

To underscore the significance of cross-dataset evaluations in real-time pain detection

[2] `https://www.scopus.com/`, Query: TITLE-ABS-KEY ((face OR facial) AND pain AND (detect* OR recogni* OR predict* OR "intensity estimation" OR classif*) AND (cross-dataset OR cross-database OR cross-corpus OR generaliza* OR transferab*))

applications, Dai et al. [2019] conducted a comparative study of models trained on AffectNet emotion and UNBC pain datasets using extracted AUs and RGB face images. They initially evaluated these models in real-time scenarios involving posed facial expressions. Their findings revealed that CNNs trained on RGB images exhibited a tendency to classify all posed expressions, including pain, as no-pain. This observation prompted the authors to suggest that these CNNs might be learning dataset-specific features. In contrast, the AU-based model demonstrated promising performance in the real-time test. However, subsequent evaluation with images from the BioVid heat pain dataset revealed a substantial decline in the model's performance. A manual examination of randomly selected instances from the BioVid dataset indicated that, in contrast to the UNBC dataset, participants in the BioVid dataset frequently closed their eyes even during non-pain conditions.

Tavakolian et al. [2020] trained spatio-temporal models on the UNBC and BioVid datasets and evaluated their performance across these datasets. Cross-dataset evaluations revealed a decline in performance for both models, with the UNBC model experiencing a more significant drop. Furthermore, they conducted additional evaluations incorporating a portion of the test dataset (10 - 50%) into the training process. This approach consistently led to an improvement in cross-dataset performance.

Rezaei et al. [2020] employed a contrastive training approach to train pain estimation models on the UNBC shoulder pain and UofR dementia pain [Hadjistavropoulos et al., 2018] datasets. Their cross-dataset evaluations revealed a notable drop in the performance of the UNBC model when applied to the dementia pain dataset. In contrast, the dementia pain model exhibited a surprisingly higher performance on the UNBC dataset compared to its within-dataset evaluation.

To investigate the impact of combining pain datasets, Zarghami et al. [2023] trained pain detection models using the UNBC dataset, a self-acquired sedation dataset, and a combined dataset. They discovered that the model trained on the sedation dataset exhibited the best performance, whereas combining datasets resulted in a performance drop. They attributed this observation to the prevalence of eye closure in the sedation dataset for both pain and no-pain classes, in contrast to the UNBC dataset, where closed eyes are strongly correlated with pain. Notably, the authors did not evaluate their models on the UNBC dataset, preventing a comprehensive cross-dataset comparison.

The UNBC shoulder pain dataset is widely employed in cross-dataset studies. However, models trained on UNBC often exhibit poor performance when applied to other datasets. While the reasons for this performance gap have not been thoroughly investigated, Dai et al. and Zarghami et al. conducted manual inspections of the datasets to propose potential contributing factors for the performance discrepancy.

4.2.4 Research Gaps

The literature review aimed to gain insights into the approaches employed by previous studies to assess the learned features and generalization capabilities of pain detection models. The review identified three key research gaps: the lack of quantitative analysis to interpret XAI visualizations, the absence of systematic XAI approaches for comparing models trained on different datasets, and the limited investigation into the performance drop ob-

Paper	Pain Datasets	Cross-dataset Outcome	Outcome Analysis
Othman et al. [2019]	D1: BioVid D2: X-ITE	D1 model on D2: ∼ D2 model on D1: ∼	–
Dai et al. [2019]	D1: UNBC D2: Own D3: BioVid	D1 model on D2: ∼ D1 model on D3: ↓	Manual
Tavakolian et al. [2020]	D1: UNBC D2: BioVid	D1 model on D2: ↓ D2 model on D1: ∼	–
Rezaei et al. [2020]	D1: UNBC D2: Dementia	D1 model on D2: ↓ D2 model on D1: ↑	–
Zarghami et al. [2023]	D1: UNBC D2: Own	All combinations were not tested	Manual
Prajod et al. [2022a]*	D1: UNBC D2: BioVid	D1 model on D2: ↓ D2 model on D1: ∼	**XAI**

*Table 4.4: An overview of the existing works that perform cross-dataset evaluations to assess the generalizability of their pain models. The downward arrow (↓) indicates a decline in cross-dataset performance compared to the model's within-dataset performance, the upward arrow (↑) signifies an improvement in cross-dataset performance, and the tilde symbol (∼) represents comparable within- and cross-dataset performances. Performance differences below 5% are considered similar. The entries marked with * are expanded in the subsequent sections of this chapter.*

served in cross-dataset evaluations. These gaps are addressed in the subsequent sections of this chapter.

- **Quantitative analysis of XAI heatmaps**: The studies reviewed in Table 4.2 primarily rely on manual inspections of heatmaps, lacking a quantitative approach to analyzing these visualizations. A notable exception is the work by Rieger et al. [2020], which introduces an initial step towards quantitative analysis by quantifying the instances where highlighted pixels lie outside pre-defined bounding boxes. However, the outcomes of the manual inspections are often subjective and it remains unclear whether these findings represent statistically significant patterns. Hence, incorporating statistical testing into manual inspection procedures is crucial to determine whether the observed model behavior differs significantly across pain classes or datasets.

- **XAI for systematic comparison of models**: A notable gap in the literature is the absence of comparative studies that analyze and contrast explanations generated by models trained on different pain datasets. Existing research (see Table 4.3) primarily focuses on generating explanations for either a single model or multiple models trained on the same dataset. This lack of comparative analysis hinders the understanding of the variations in learned feature representations among pain models trained on different datasets. Hence, there's a need to develop a systematic XAI procedure that facilitates a visual and statistical comparison of differences between pain models trained on different datasets.

- **Augmenting cross-dataset evaluations with XAI**: Cross-dataset evaluations to assess the generalizability of pain detection models are surprisingly uncommon. Even fewer studies delve beyond performance evaluations to identify the underlying causes of cross-dataset disparities. As seen from Table 4.4, existing studies rely on subjective manual inspections of datasets, which are time-consuming and lack statistical rigor. This chapter presents a systematic XAI-based comparison approach to identify the differences between pain models that contribute to cross-dataset performance gaps.

4.3 Transfer Learning Pain from Emotions

State-of-the-art approaches for facial expression recognition predominantly employ deep learning techniques, which can learn task-specific representations from raw data inputs [Luqin, 2019; Li and Deng, 2020; Gkikas and Tsiknakis, 2023a]. While these methods consistently outperform traditional handcrafted features in terms of classification performance, they come with a substantial requirement for large amounts of annotated training data. This requirement poses a significant challenge in sensitive classification tasks, such as automatic pain detection, where data scarcity is prevalent due to limited patient contact, privacy concerns, and strict adherence to ethical guidelines [Kunz et al., 2017; Cowie et al., 2017; Charlton, 1995]. Hence, pain datasets are typically limited in size, with only a few pain samples available for analysis [Wang et al., 2017; Hassan et al., 2019; Xiang et al., 2022].

To address the data scarcity challenge, a frequently employed technique is transfer learning [Wang et al., 2017; Coutrin et al., 2022]. This technique involves re-using certain parameters from a pre-trained model while training the remaining parameters on a smaller target dataset. Ideally, the pre-trained model should originate from a domain closely resembling the target domain [Yosinski et al., 2014; Weiss et al., 2016]. In the case of pain detection, emotion recognition can be leveraged as a related task, given that both pain and emotion expressions can be characterized by facial AUs. Facial AUs represent a collection of distinct facial muscle movements that correspond to a facial expression. Notably, pain and Ekman's basic emotions have some overlapping AUs [Kappesser and de Williams, 2002; Simon et al., 2008; Kunz et al., 2019]. Moreover, previous work by Florea et al. [2015] showed that data representations of hand-crafted features learned in emotion recognition task can be leveraged for pain estimation. Additionally, emotion recognition from facial expressions is a well-explored task with existing large datasets such as AffectNet [Mollahosseini et al., 2017] and SEWA DB [Kossaifi et al., 2019], making it a promising source task. All the pain detection models discussed in this chapter were trained using a transfer learning approach, with the emotion recognition model as the source model.

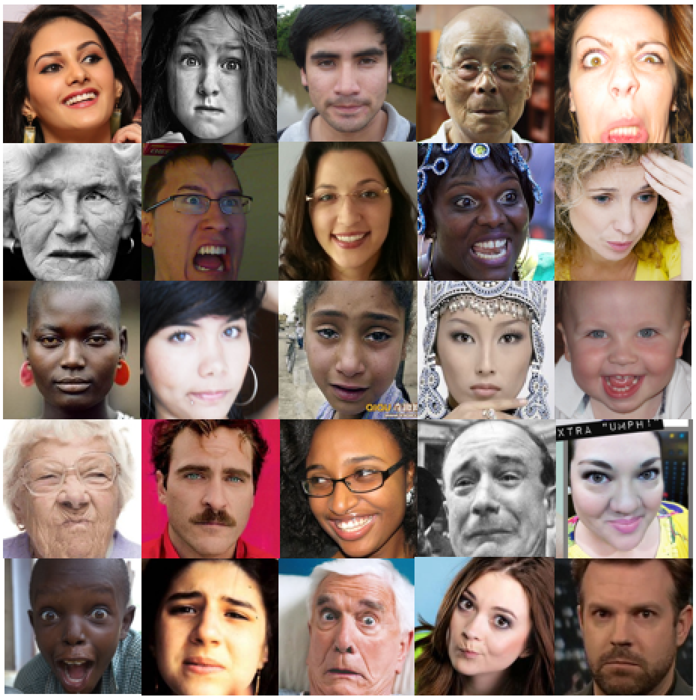


Figure 4.2: Sample images from the AffectNet dataset comprising images collected from the internet. The copyright remains with the dataset creators [Mollahosseini et al., 2017].

4.3.1 Emotion Recognition Model

Dataset

The facial emotion recognition model was trained using the AffectNet dataset [Mollahosseini et al., 2017]. The dataset contains over 400,000 annotated facial images distributed across 11 classes - Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, Contempt, None, Uncertain, and Non-Face. The images were collected from the internet through multilingual search queries containing emotional keywords. Some example images are shown in Figure 4.2. Each image was assigned a discrete emotion class through a manual annotation process involving 12 experts. Additionally, annotations were extended to include valence and arousal values. Notably, the dataset is imbalanced, with categories like 'Happy' and 'Neutral' containing a substantial number of samples (over 75,000), while 'Disgust' and 'Contempt' have fewer samples (around 4,000). Images belonging to 'None', 'Uncertain', and 'Non-face' were excluded due to the absence of a relevant emotion label. Around 90% of the valid images were used for training and the remaining 10% for validation. The performance of the model was evaluated on the test set provided by Mollahosseini et al. as part of the dataset. This test set contains a total of 4000 images (500 per class × 8 classes).

Training Procedure

The emotion recognition model was trained by leveraging the VGG16 architecture [Simonyan and Zisserman, 2014], which contains five convolution blocks. Although AffectNet is a large dataset with sufficient images to train a deep learning model from scratch, the VGG16 network is pre-trained on ImageNet dataset [Russakovsky et al., 2015] to enhance

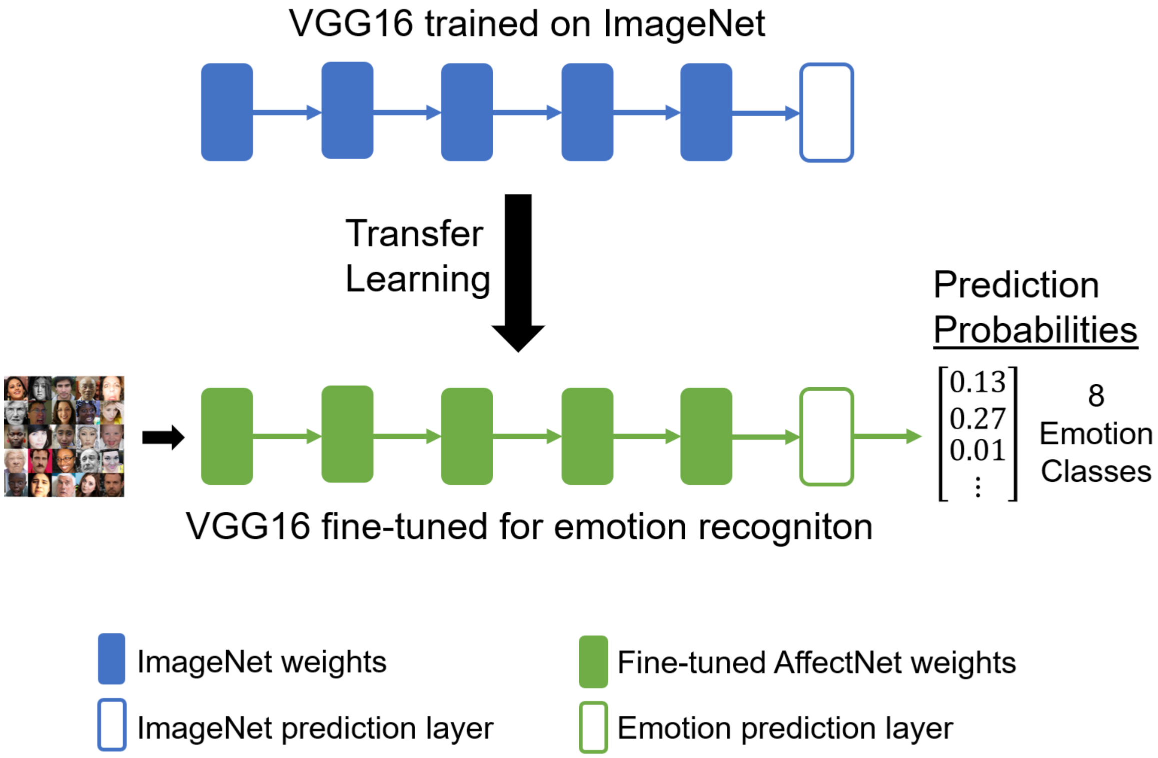

VGG16 trained on ImageNet
Transfer Learning
Prediction Probabilities
0.13
0.27
0.01
⋮
8 Emotion Classes
VGG16 fine-tuned for emotion recogniton
ImageNet weights
Fine-tuned AffectNet weights
ImageNet prediction layer
Emotion prediction layer

Figure 4.3: Visualization of the training process employed for the emotion recognition model. The model is trained on the AffectNet dataset to output prediction probabilities for eight emotion classes.

its performance and training efficiency [Yen and Li, 2022]. This network is connected to a fully connected layer, followed by a dense layer with softmax activation to predict the probability of an image belonging to each of the eight emotion classes - Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, and Contempt. As illustrated in Figure 4.3, the model underwent a full model fine-tuning, i.e., all the layers were trained on the AffectNet dataset. All input images were scaled to default VGG16 dimensions (224 × 224). During training, the images were randomly flipped horizontally to increase variations in the input. The training process employed the Stochastic Gradient Descent (SGD) optimizer (learning rate = 0.01) and a weighted focal loss function. The focal loss function [Lin et al., 2017] is computed as follows:

$$focal\_loss = (1 - p_t)^{\gamma} \times cross\_entropy\_loss \tag{4.1}$$

The variable p_t represents the predicted probability of a sample belonging to its true class (t), and γ is a hyperparameter that was empirically set to 5. Given the substantial imbalance in the AffectNet dataset, a weighted focal loss function was implemented using the weighting scheme proposed by Cui et al. [2019], which is given by:

$$weighted\_loss = \frac{1 - \beta}{1 - \beta^{samples\_per\_cls}} \times focal\_loss \tag{4.2}$$

Following the examples provided by Cui et al., the hyperparameter β was empirically set to 0.99998.

Emotion	Precision	Recall	F1-score
Neutral	0.45	0.56	0.50
Happy	0.62	0.81	0.70
Sad	0.64	0.54	0.58
Surprise	0.58	0.52	0.55
Fear	0.65	0.64	0.65
Disgust	0.57	0.61	0.59
Anger	0.55	0.54	0.54
Contempt	0.59	0.41	0.48
Average	0.58	0.58	0.57

Table 4.5: Performance of the emotion recognition model in terms of precision, recall, and F1-score for the eight emotion classes

Evaluation

The evaluation of the emotion recognition model was conducted on the dedicated test set within the AffectNet dataset, consisting of 4000 images evenly distributed among the eight classes. Class-wise precision, recall, and F1-score were computed to evaluate the model's performance on individual emotions. Additionally, average F1-score and accuracy were calculated as measures of overall performance. The class-wise prediction performance is presented in Table 4.5, with the model achieving an accuracy of 58% and an average F1-score of 57%. The overall performance aligns with the baseline performance reported for this test set [Mahoor, 2017].

4.3.2 Transfer Learning Pain

There are two primary methods for transfer learning feature representations in deep neural networks: freezing and fine-tuning. In this chapter, a hybrid approach is adopted, wherein all layer weights from the pre-trained model are transferred to the target model, and a subset of layers is frozen, while the remaining layers undergo fine-tuning on the target dataset. In the hybrid approach, no layer is randomly initialization. The three approaches are visualized in Figure 4.4.

The following paragraphs describe the transfer learning process for Section 4.4. The pain transfer learning process for Section 4.5 is a specific instance of the process followed in Section 4.4, where no layer is frozen, i.e., the whole pre-trained model undergoes fine-tuning.

Reduced UNBC Pain Dataset

For training pain detection models, the facial images from the UNBC-McMaster Shoulder Pain Expression Database [Lucey et al., 2011] are utilized. This dataset consists of image sequences derived from video recordings of 25 participants experiencing shoulder pain;

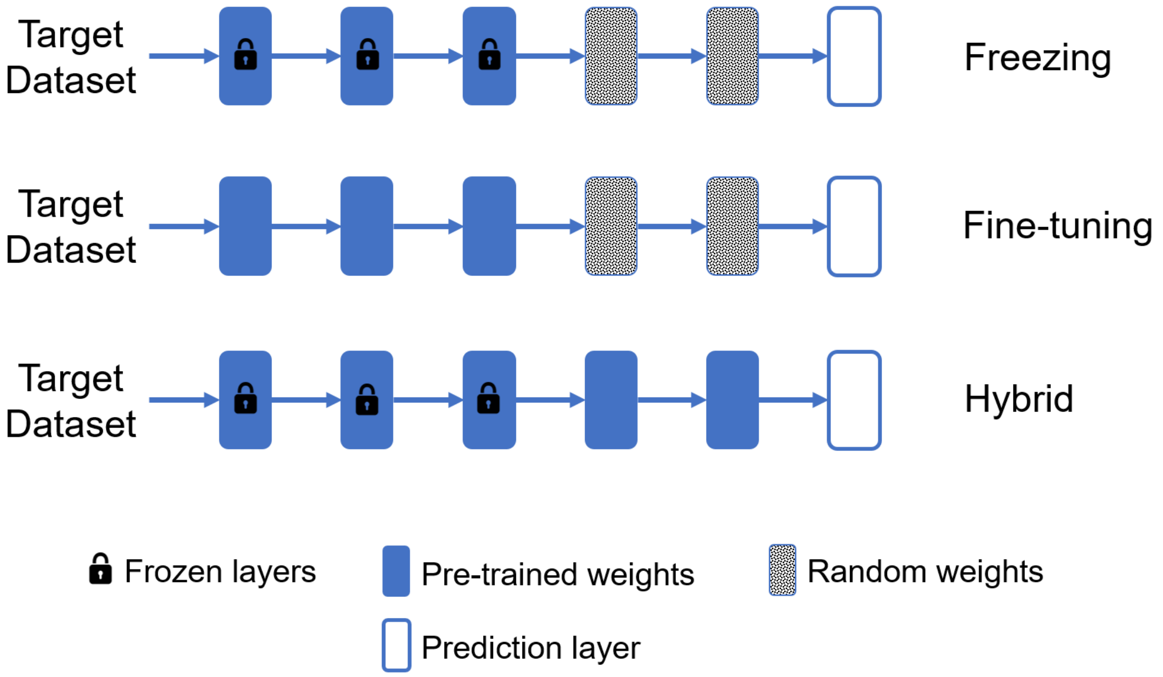


Figure 4.4: Depiction of the freezing, fine-tuning, and hybrid approaches for transfer learning. The hybrid approach utilizes both freezing and fine-tuning to train a model.

however, only four participants consented to the usage of their images in publications. Participants were instructed to perform specific arm movements, serving as stimuli for pain expression. Each image in the dataset is annotated with a PSPI score on a scale from 0 (no pain) to 15 (extreme pain). In this chapter, images with a PSPI score of 0 are categorized as no-pain images, while those with scores greater than 0 are considered pain images. The dataset comprises a total of 48,398 images, with 40,029 classified as no-pain and 8,369 as pain images. Due to the visual analysis in this chapter, the test set exclusively includes images from the four participants who consented to publication. Some examples from the test set are presented in Figure 4.5.

Since these images originate from video sequences, many frames are similar. To mitigate redundancy in the training and validation sets, images are randomly chosen, adhering to the following conditions:

(a) There is at least one pain and one no-pain image from each of the 21 participants

(b) At least five images separate two selected images from the same sequence

Additionally, to address dataset imbalance, a deliberate effort is made to balance the dataset. Following the criteria to mitigate redundancy, 1,000 images (500 pain and 500 no-pain) are randomly selected. Following a 90-10 split, this reduced dataset is divided into training (900 images) and validation (100 images) sets.

Pain Training

The emotion recognition model trained on AffectNet was leveraged as the pre-trained model for transfer learning pain detection. As shown in Figure 4.6, the transfer learning

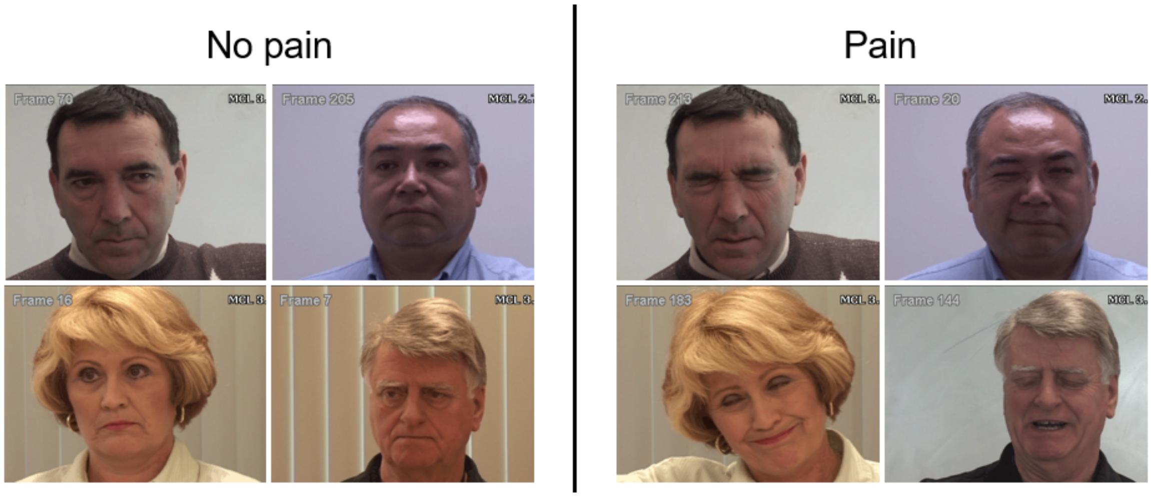


Figure 4.5: Some examples of pain and no-pain images from the UNBC-McMaster shoulder pain dataset. The copyright remains with the dataset creators [Lucey et al., 2011].

process involved freezing and fine-tuning of layers. The VGG16 architecture is composed of five convolutional blocks, with each block comprising several convolutional layers and ending with a pooling layer. The weights of all convolution blocks of the emotion recognition model were copied, and the initial blocks were frozen, rendering them unchangeable by pain training. For the analyses detailed in Section 4.4, the number of frozen blocks was systematically varied from 0 (allowing all blocks for pain training) to 5 (excluding all blocks from pain training, with only the prediction layers trainable). This method resulted in six distinct pain detection models. For ease of referencing, the models are denoted as 'FrozenBlocks$\langle i \rangle$', where i ranges from 0 to 5, indicating the number of frozen convolutional blocks during pain training. Each FrozenBlocks$\langle i \rangle$ model was trained by freezing the initial i convolutional blocks and fine-tuning the remaining blocks using the pain dataset. Specifically, FrozenBlocks5 represents a model with all convolutional blocks frozen, and only the output layers were trained to detect pain. Conversely, FrozenBlocks0 corresponds to a model where no blocks were frozen, signifying that all layers underwent fine-tuning with the pain dataset.

All the models were fine-tuned using the aforementioned reduced UNBC pain dataset. The prediction layer employed softmax activation to output whether a facial image expressed pain or not. The input images were scaled to default VGG16 dimension. Due to the small size of the dataset, image augmentation was employed during training using the Keras data augmentation options, including rotation (within ± 25°), height shift (within ± 10%), width shift (within ± 10%), shear (within ± 10%), zoom (within ± 10%), and horizontal flip. Unlike the AffectNet dataset, the reduced UNBC pain dataset was intentionally balanced, allowing the use of unweighted focal loss (Equation 4.1). The hyperparameter γ was empirically set to 2, reflecting a balance between focusing on difficult-to-classify samples and general model performance.

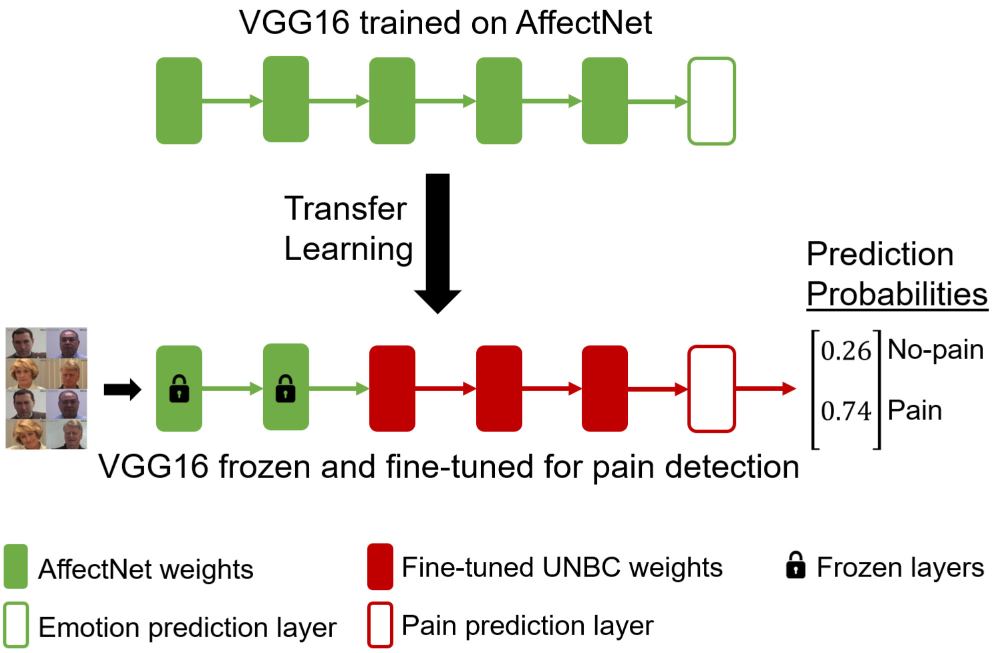


Figure 4.6: Illustration of the process followed for transfer learning pain detection from an emotion recognition model. The image shows the process of training a pain detection model by freezing the first two blocks of the emotion recognition model.

4.3.3 Evaluation

The models were evaluated on a test set derived from the UNBC dataset, consisting of images from four participants. During testing, all images from these four participants were utilized instead of a reduced version, resulting in an imbalanced set. Consequently, the macro-averages of the performance metrics, which compute the metric for each class and average them, were calculated to ensure equal consideration for every class. The performance metrics, including recall, F1-score, and accuracy, for the six pain detection models are presented in Table 4.6.

While the recall for the no-pain class was the highest in FrozenBlocks4 and Frozen-

Models	Recall			F1-score			Accuracy
	No-pain	Pain	Avg.	No-pain	Pain	Avg.	
FrozenB5	**0.99**	0.54	0.76	0.95	0.66	0.81	0.92
FrozenB4	**0.99**	0.56	0.78	**0.96**	0.69	0.83	0.92
FrozenB3	0.98	0.69	0.83	**0.96**	**0.76**	**0.86**	**0.93**
FrozenB2	0.96	**0.71**	**0.84**	0.95	0.73	0.84	0.92
FrozenB1	0.95	**0.71**	0.83	0.95	0.72	0.83	0.91
FrozenB0	0.95	**0.71**	0.83	0.95	0.71	0.83	0.91

Table 4.6: Recall and F1-score for no-pain and pain classes, along with the macro averages and accuracy for the six pain detection models. The model names are abbreviated as FrozenB⟨i⟩

Blocks5, it consistently remained high (≥ 95%). This could be attributed to the similarity between the no-pain class and the neutral class in emotion recognition, suggesting pre-existing discerning features of this class. As anticipated, models with a higher number of available convolution blocks for pain training (FrozenBlocks2, FrozenBlocks1, FrozenBlocks0) achieved higher pain recall. Notably, the pain recall saturated beyond a certain point, specifically, unfreezing the initial two blocks for pain training did not improve the recall. The best overall performance, as measured by accuracy and F1-score, was achieved by FrozenBlocks3. This observation suggests a potential risk of over-tuning models on small datasets with fewer frozen blocks.

Paper	Accuracy	F1-score
Sikka et al. [2014]	0.84	0.52
Ruiz et al. [2014]	0.86	-
Pedersen [2015]	0.86	-
Wu et al. [2015]	0.85	0.78
Shrivastava et al. [2015]	0.88	-
Chen et al. [2015]	0.87	-
Yang et al. [2016]	0.73	-
Roy et al. [2016]	0.88	-
Guo et al. [2016]	0.85	-
Rathee and Ganotra [2016]	0.90	-
Kharghanian et al. [2016]	0.87	0.86
Chen et al. [2017]	0.91	0.54
Rodriguez et al. [2017]	0.84	-
Kumawat et al. [2019]	0.87	-
Abedi et al. [2020]	0.73	-
Semwal and Londhe [2021a]	0.67	0.62
Reichard et al. [2022]	0.81	0.48
This chapter (FrozenBlocks3)	0.93	0.86

Table 4.7: Performance (accuracy and F1-score) of UNBC shoulder pain detection models from the literature

Table 4.7 summarizes the performance of pain detection models from the literature, providing a benchmark for the models presented in this section. The performances were obtained through a literature search in the Scopus[3] database. Only binary classifiers trained on the UNBC shoulder pain dataset and evaluated on unseen participants (leave-one-subject-out, participant hold-out test set, etc.) were considered for comparison. The models presented in this section outperform the existing models in terms of accuracy.

[3] `https://www.scopus.com/`, Query: TITLE-ABS-KEY (pain AND (detect* OR recogni* OR predict* OR estimat* OR classif* OR learn*) AND unbc)

4.4 An XAI-based Approach to Assessing Forgetting

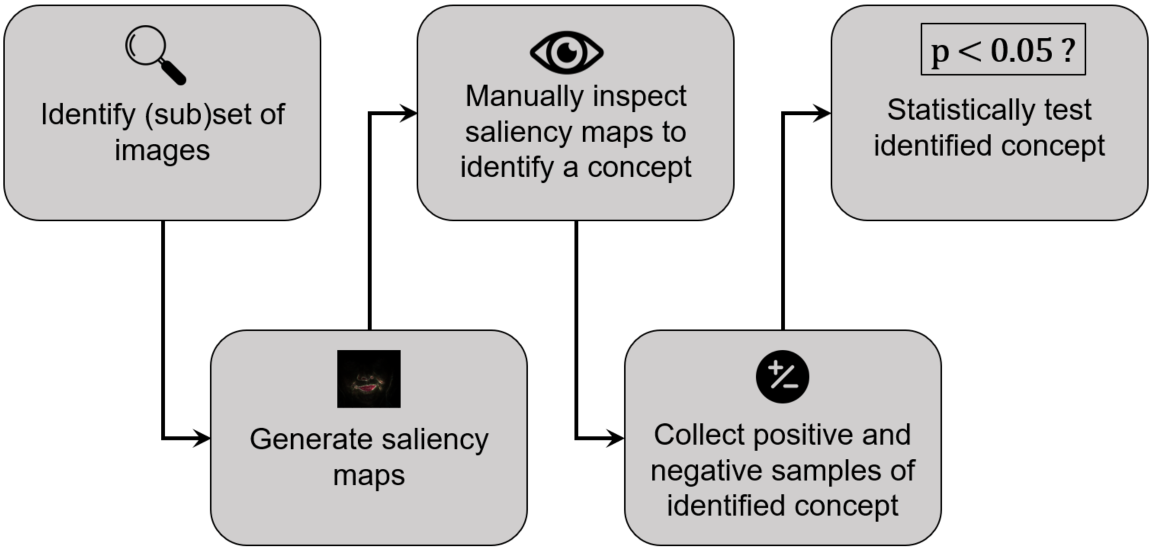

Identify (sub)set of images
Generate saliency maps
Manually inspect saliency maps to identify a concept
Collect positive and negative samples of identified concept
p < 0.05 ?
Statistically test identified concept

Figure 4.7: Visualization of the XAI-based assessment process implemented in this chapter.

The fine-tuning approach of transfer learning often leads to considerable changes to the pre-trained model weights and thus alters the internal feature representations. This phenomenon is called Catastrophic Forgetting, as the new model 'forgets' the features required to perform well on the original task Kemker et al. [2018]. The work by Khorrami et al. [2015] involved analyzing the learned representations of an emotion recognition model, revealing its capability to learn AUs as features. In the case of transfer learning pain, when more unfrozen layers are available for training pain, the model becomes increasingly tailored for the dataset. It is crucial to examine the learned representations of the model to ensure that it has not overlooked certain AUs, leading to forgetting AUs relevant to pain detection. This oversight might occur due to the dataset lacking ample samples of specific well-known pain patterns, making it less suitable for training a generic pain detection model.

To assess the learned representations and identify any forgotten features, a multi-step process was devised and implemented. This process is visualized in Figure 4.7. First, the pain detection models were evaluated on the emotion recognition task, focusing on class-wise recall. The model was deemed to have forgotten the recognition of a specific emotion if there was a significant decrease in recall after the transfer learning process. At the end of the first step, the images contributing to the decline in recall were identified. Second, these images were used to generate visualizations of learned features using XAI techniques. In the third step, the generated visualizations were manually examined to pinpoint the forgotten AUs. Finally, the learned representations of the models were evaluated for the forgotten AUs to determine whether the differences were statistically significant. This process helps in identifying AUs forgotten during pain transfer learning, allowing for a comparison with AUs typically associated with pain expression.

4.4.1 Re-train to Identify Forgetting

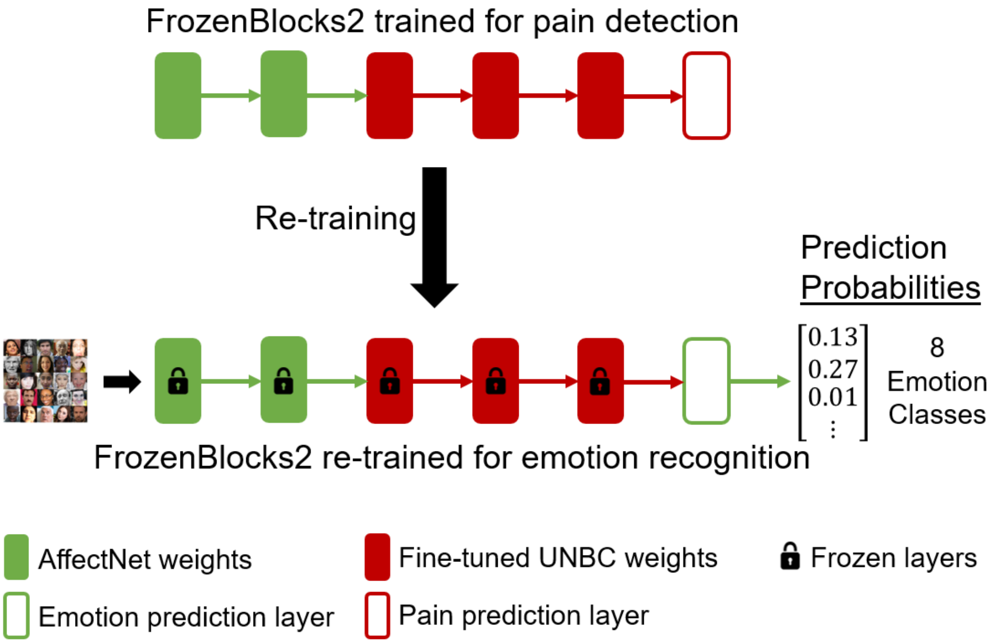


Figure 4.8: Illustration of the process used to re-train the pain models for emotion recognition. In this depiction, all convolutional blocks of the pain model were frozen, and only the prediction layers were trained on the AffectNet dataset. The image illustrates the learned weights of FrozenBlocks2 as a representative example.

Depending on the number of fine-tuned layers, the hybrid transfer learning approach can yield many different pain models and associated feature representations. In this chapter, the convolutional blocks were frozen rather than individual layers to limit the number of models considered. There are five convolutional blocks in VGG16 architecture, and varying the number of frozen blocks during transfer learning from 0 to 5 yielded six different pain detection models.

The first step in assessing forgotten features involves evaluating how these models perform on emotion recognition after adapting feature representations for pain detection. This is accomplished through a re-training methodology, as illustrated in Figure 4.8. The objective is to evaluate the capability of pain detection models to still recognize emotions using the AffectNet test set. A decrease in the recall of a specific emotion suggests potential forgetting of AUs crucial for recognizing that emotion. While the models are expected to forget AUs irrelevant to pain expression, there is a possibility of forgetting AUs that play a role in pain detection, contingent on their representation in the pain dataset. For instance, Jaw Drop (AU 26) is an AU that may manifest in both pain and surprise expressions. If the pain training dataset lacks sufficient images with AU 26, the model may forget this AU learned for surprise, despite its relevance to pain expression.

As part of the re-training method, all convolutional blocks of a pain model were frozen to preserve the learned feature representations. The output layers were then modified to predict the eight emotion classes and subsequently trained on the AffectNet dataset. The re-training process utilized the same optimizer, data augmentation techniques, loss

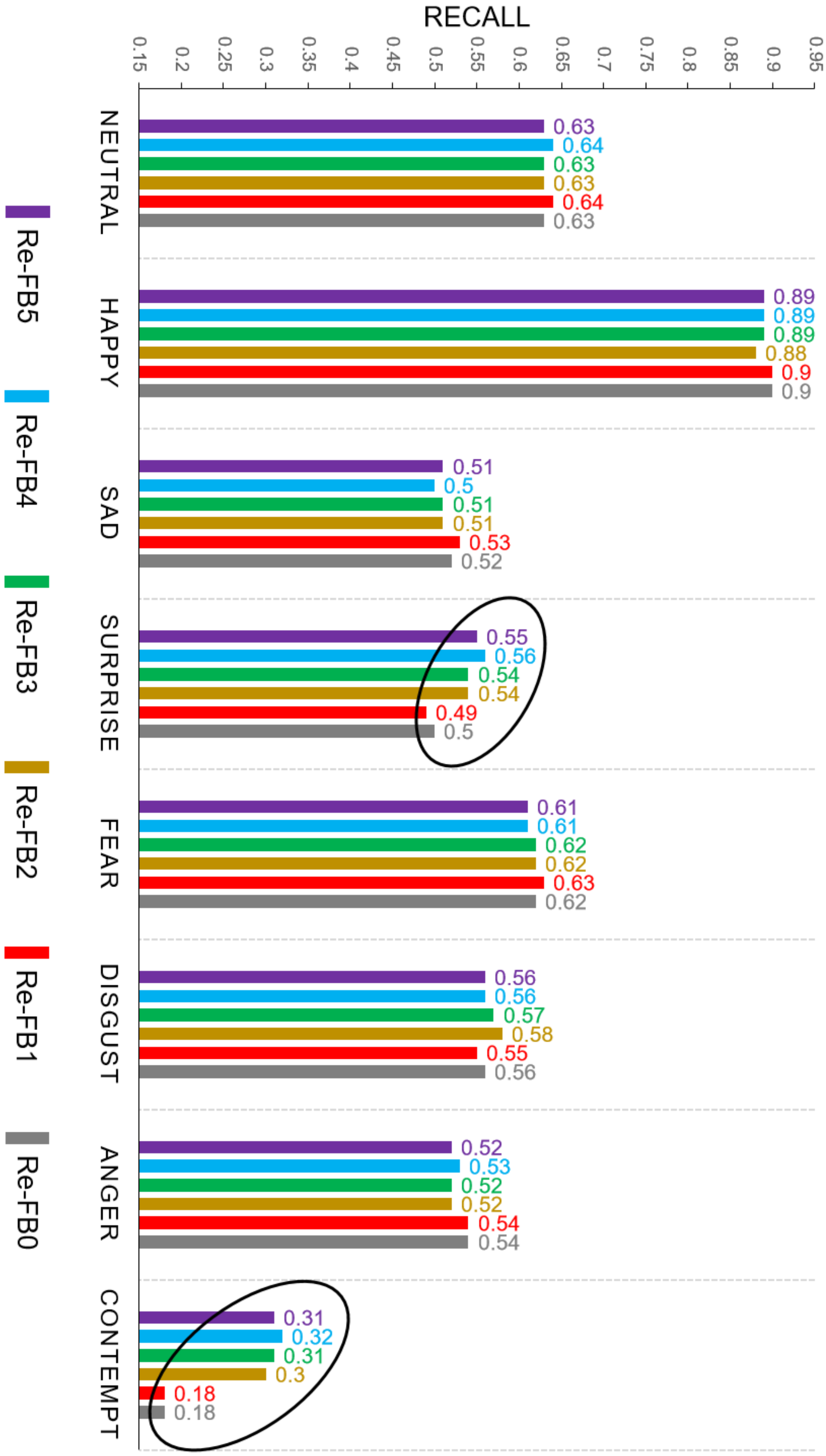


Figure 4.9: Class-wise recalls of the six pain models obtained by re-training their output layers for emotion recognition. The black ellipses highlight the decrease in recall for the Surprise and Contempt classes. The copyright remains with the authors [Prajod et al., 2022b].

function, and hyperparameters as previously described in Section 4.3.1. This re-training process was followed for all six pain models, and the resulting models were compared in terms of their recalls for each emotion. Following a comparison strategy similar to Dietterich [1998], McNemar's test was employed to determine if the differences in recalls were significant.

Given that the weights of the convolutional blocks in the re-trained models are identical to the pain models, these models are denoted as Re-trained FrozenBlocks$\langle i \rangle$ ($i \in \{0, 1, 2, 3, 4, 5\}$), or Re-FB$\langle i \rangle$ for short. Figure 4.9 shows the recalls for the eight emotions by utilizing the six re-trained models. The recalls for Surprise and Contempt classes decreased with an increase in pain training blocks, dropping notably for Re-FB1 and Re-FB0. To further investigate this decline in recall, a detailed analysis was conducted by comparing Re-FB5 and Re-FB0. McNemar's test between the two models revealed a significant difference in recall for both Surprise (p-value: 1.56×10^{-4}) and Contempt (p-value: 8.19×10^{-18}) classes.

4.4.2 Visual Analysis

While analyzing the class-wise recalls of the re-trained models can highlight the specific emotions affected by pain training, it doesn't provide insight into the learned or forgotten features. To delve into these forgotten features or concepts, the learned representations of Re-FB5 and Re-FB0 were compared. Re-FB5 mirrors the original emotion recognition model, while Re-FB0 underwent the most substantial modifications due to pain training. To facilitate this comparison, the images causing the drop in recalls for the Surprise and Contempt classes were identified. In other words, these were the AffectNet test set images belonging to the Surprise and Contempt classes, which were correctly predicted by Re-FB5 but were incorrectly predicted by Re-FB0.

To visualize the learned feature representations of both models, a saliency map was generated for each identified image. A saliency map highlights the pixels of an input image that have the most influence on the model's prediction probability for a specific class. The saliency maps were created using the XAI technique called LRP Bach et al. [2015]. Following the recommendations by Montavon et al. [2019] and Sixt et al. [2020], the chosen LRP variant used the z-rule for fully connected layers and the z^+-rule for convolution layers. The saliency maps were generated using the iNNvestigate library [Alber et al., 2019]. Additionally, the saliency maps were normalized between 0 and 1, where a higher value indicates higher relevance to the prediction.

To translate relevant image areas into semantically meaningful concepts, the interpretation needs to be done visually by a human. The underlying intuition is that the visually identified concepts helped Re-FB5 to correctly predict the emotion but were forgotten by Re-FB0. However, it is crucial to consider the sensitivity of neural networks to small differences. Since both models were fine-tuned based on the same emotion recognition weights, the saliency maps for the same input image often appear indistinguishable to the human eye (see figure 4.10). To make those differences visible, new saliency maps were generated by subtracting the raw saliency maps from each other and again normalizing these differences between 0 and 1.

As observed in Figure 4.9 and subsequent McNemar's tests, there is a significant drop

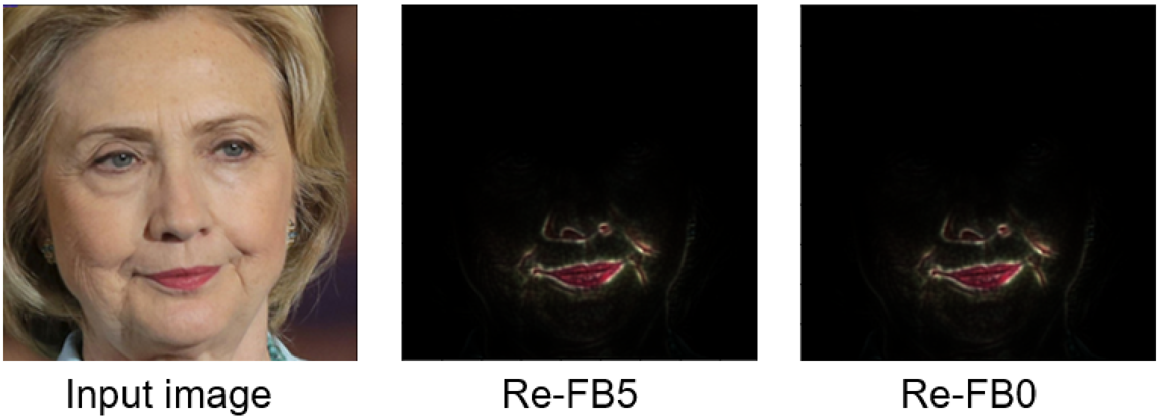


Figure 4.10: The image on the left is an example from the Contempt class, correctly predicted by Re-FB5 but not by Re-FB0. The middle and right images illustrate the saliency maps corresponding to the Contempt class for Re-FB5 and Re-FB0, respectively. The copyright remains with the authors [Prajod et al., 2022b]

in recalls for two emotions - Surprise and Contempt. Previous work by Khorrami et al. [2015], which examined the learned representations in deep neural networks for emotion recognition, revealed that these networks learn facial AUs. Therefore, when inspecting the saliency maps to identify forgotten features, AUs were specifically considered as potential learned concepts.

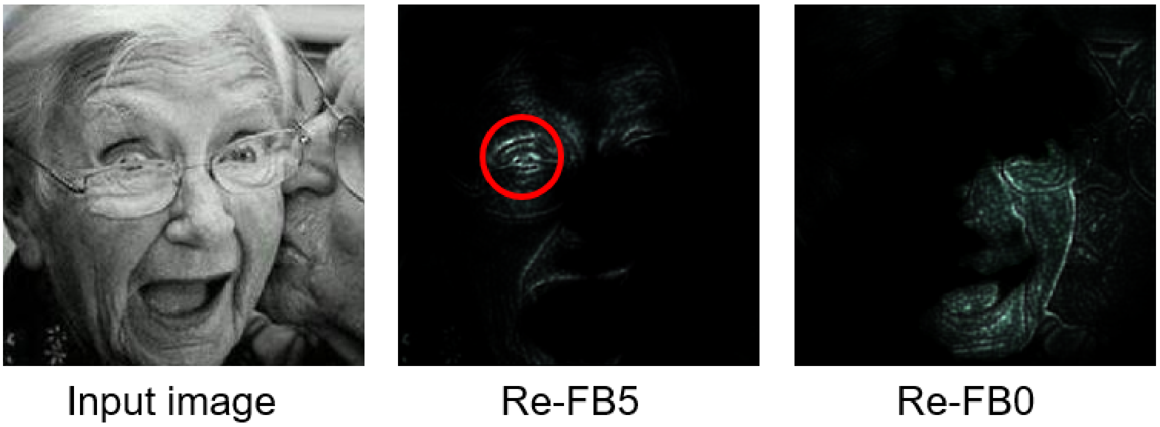


Figure 4.11: An illustrative comparison of saliency maps for a Surprise image, highlighting differences in the learned representations of Re-FB5 and Re-FB0. The red circle highlights the pixels that Re-FB5 focused on, which is indicative of AU5. The copyright remains with the authors [Prajod et al., 2022b].

A visual analysis of the saliency maps for test images from the Surprise class revealed that, in comparison to Re-FB0, Re-FB5 focused more on AU5 (upper lid raise). This emphasis on AU5 is depicted in Figure 4.11. Similarly, as illustrated in Figure 4.12, saliency maps for Contempt images revealed a greater emphasis on dimples (corresponding to AU14) in Re-FB5 compared to Re-FB0. These observations suggest that, although both models utilized AU5 and AU14 to some extent, Re-FB5 exhibited a more pronounced representation of these AUs. In other words, allocating more blocks for pain training resulted in a reduced

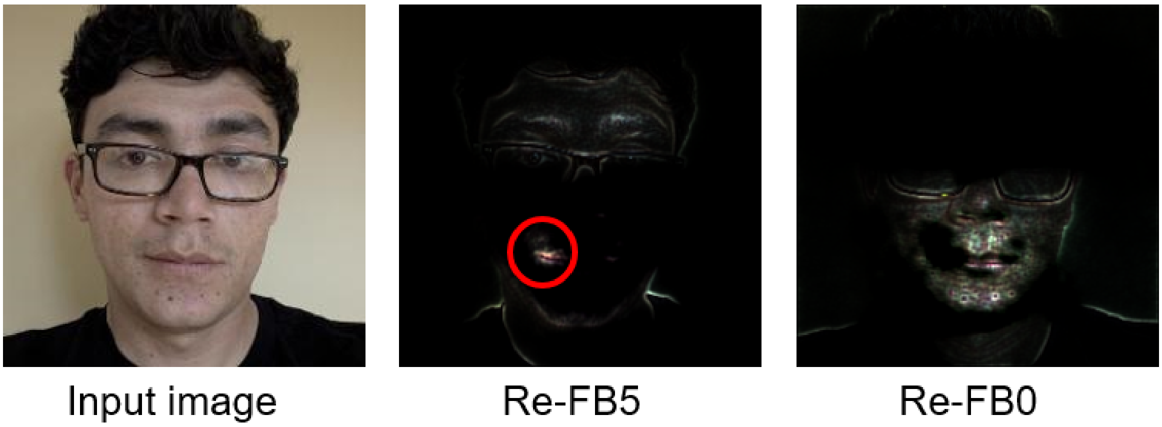


Figure 4.12: An illustration comparing the saliency maps for a Contempt image. The image highlights differences in the learned representations of Re-FB5 and Re-FB0. The red circle denotes a facial region that Re-FB5 focused on, which can be linked to AU14. The copyright remains with the authors [Prajod et al., 2022b].

emphasis on AU5 and AU14.

4.4.3 Embedded Concept Detection

After identifying the forgotten concepts — specifically, the AUs influenced by pain training — the next step involves a statistical validation of this observation. An embedded concept detection technique [Kim et al., 2018a] was employed to assess potential statistical differences in the learned representations of the models concerning AU5 and AU14. This technique involved training a binary linear classifier for concept detection, utilizing the output of an intermediate layer in the network to generate feature vectors for input images. In other words, the intermediate layer of the neural network acted as a feature extractor for the concept detection classifier. The efficacy of the classifier in concept detection indicates the extent to which the network has learned the specific concept. Additionally, a statistical comparison of the performances of two classifiers (derived from two distinct models) can be used to determine whether one representation embedded the concept more effectively than the other.

While the identified concepts were AUs, the AffectNet dataset lacked AU annotations. According to Kim et al., the concept detection classifier need not be trained on the same dataset as the neural network. Hence, the concept classifiers were trained on another facial emotion expression dataset, namely, the CK+ dataset [Lucey et al., 2010]. This is a relatively smaller dataset that contains images of acted emotions with manually annotated AU labels. Given that CK+ images with AU5 often exhibit a wide-open mouth corresponding to AU26 (jaw drop), the mouth area (bottom of the images) was cropped to ensure that AU5 classifiers were trained exclusively on AU5 and not influenced by AU26. This cropping was specifically applied to AU5 classifiers and not AU14. As noted by Kim et al., the cropped versions of a concept image do not impede concept detection.

The output of the last convolution block of Re-FB5 and Re-FB0 was used to generate two sets of feature vectors for the input images. For each set, a linear Support Vector

Machine (SVM) was trained to detect the concept (AU5 or AU14) using two-fold cross-validation. The average F1-score of the two folds served as a measure of concept detection performance. This training process was iterated 500 times using random seeds for weight initialization and fold selection. This process yielded 500 performance scores for each classifier, which were used in a paired t-test comparison. The outcome of this test determined whether there was a significant difference between the performance of SVMs trained on Re-FB5 features and Re-FB0 features. Following the approach of Dietterich [1998] for a comparison metric over 5 iterations, a 5 × 2 cross-validation paired t-test was extended to 500 iterations as suggested by Kim et al. [2018a]. This comparison metric was employed by Dietterich [1998] for five iterations in a paired t-test using 5 × 2 cross-validation. It was extended to 500 iterations, as suggested by Kim et al..

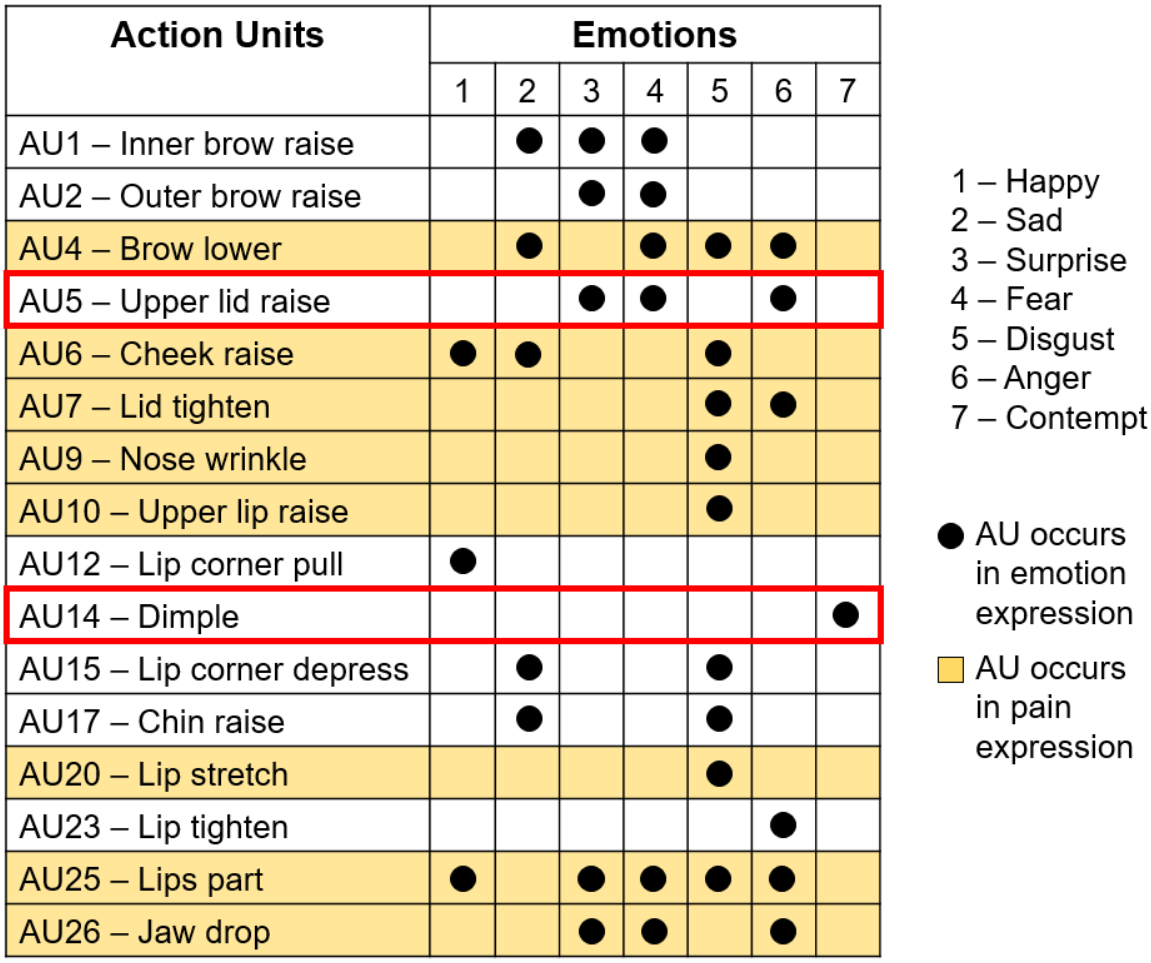
For both AU5 and AU14, the SVMs trained on Re-FB5 features significantly outperformed the classifiers trained on Re-FB0 features. In the case of AU5, Re-FB5 features achieved a mean F1-score of 84.36%, compared to 83.87% for Re-FB0 features (p-value: 3.49×10^{-19}). For AU14, Re-FB5 and Re-FB0 features achieved mean F1-scores of 74.13% and 72.41%, respectively (p-value: 4.88×10^{-34}). These results affirm the earlier observations suggesting that Re-FB0 tended to forget AU5 and AU14. However, considering that neither of the feature vectors yielded a very low F1-score, the extent of forgetting appears to be limited.

4.4.4 Insights

Both emotions and pain expressions are associated with specific characteristic AUs, and this connection serves as a foundation to investigate why certain AUs were forgotten through pain training. More importantly, it is essential to determine whether the forgotten AUs are known to occur in pain expressions. Figure 4.13 outlines the typical AUs activated during the expressions of various emotions [Simon et al., 2008; Lucey et al., 2010]. The figure also highlights the AUs activated during pain expressions [Simon et al., 2008; Kunz et al., 2019].

The activation of AU14 is essential in a typical Contempt expression, but it does not manifest in typical expressions of pain. Hence, forgetting AU14 and the subsequent decline in Contempt recall is not surprising. Similarly, AU5 is not common in typical pain expressions. However, AU5 occurs not only in typical Surprise expressions but also in typical Fear and Anger expressions. Surprisingly, forgetting AU5 did not lead to a decline in the recall of Fear and Anger classes. This observation could be explained by the lack of overlap between typical Surprise AUs in the upper face and those associated with typical pain expressions. In contrast, Fear and Anger share other eye-related AUs with pain expressions. Crucially, the forgotten AUs resulting from pain training with the UNBC dataset were not relevant for the detection of typical pain expressions.

The analyses show that the forgetting of the identified AUs, and the observed drop in recall of Surprise and Contempt are in line with the current understanding of typical pain expressions. Furthermore, these findings demonstrate the capabilities of the proposed XAI-based approach to analyze learned feature representations and assess datasets.



Action Units	Emotions						
	1	2	3	4	5	6	7
AU1 – Inner brow raise		●	●	●			
AU2 – Outer brow raise			●	●			
🟨 AU4 – Brow lower		●		●	●	●	
AU5 – Upper lid raise			●	●		●	
🟨 AU6 – Cheek raise	●	●			●		
🟨 AU7 – Lid tighten					●	●	
🟨 AU9 – Nose wrinkle					●		
🟨 AU10 – Upper lip raise					●		
AU12 – Lip corner pull	●						
AU14 – Dimple							●
AU15 – Lip corner depress		●			●		
AU17 – Chin raise		●			●		
🟨 AU20 – Lip stretch					●		
AU23 – Lip tighten						●	
🟨 AU25 – Lips part	●		●	●	●	●	
🟨 AU26 – Jaw drop			●	●		●	

1 – Happy
2 – Sad
3 – Surprise
4 – Fear
5 – Disgust
6 – Anger
7 – Contempt

● AU occurs in emotion expression

🟨 AU occurs in pain expression

Figure 4.13: List of AUs that occur in the typical expressions of various emotions and pain. The forgotten AUs (AU5 and AU14) are highlighted with red boxes. The copyright remains with the authors [Prajod et al., 2022b]

4.5 Assessing Pain Datasets

In an industrial setting, operators may encounter pain resulting from pre-existing conditions (e.g., surgery, chronic pain) or spontaneous stimuli (e.g., pressure pain). These instances of pain can be broadly categorized into clinical pain and experimental pain [Kunz et al., 2019]. Clinical pain occurs when individuals with clinically diagnosed conditions, such as those who have undergone surgery or have arthritis, experience pain while performing certain movements. On the other hand, experimental pain is induced by specific stimuli such as exposure to a high temperature or electricity. The existing pain datasets can also be classified as clinical and experimental pain datasets.

While Kunz et al. examined the differences between clinical and experimental pain in terms of manually annotated AUs, little research has been done on the differences in the learned feature representations of models trained on these datasets. This section outlines how the technique presented in Section 4.4 was employed to analyze the learned features of two models - one trained on a clinical pain dataset and the other on an experimental pain dataset. The analysis process followed the steps visualized in Figure 4.7. The cross-dataset evaluations of the two models formed the basis for identifying the initial set of images. This analysis provides insights into the robustness of each model and the extent to which they can be applied to an industrial scenario for generic pain detection.

4.5.1 Clinical and Experimental Pain Datasets

This analysis utilized two existing datasets - the UNBC-McMaster Shoulder Pain Dataset and the BioVid Heat Pain Dataset. The UNBC shoulder pain dataset serves as a representative clinical pain dataset, comprising images of individuals who underwent surgery while performing movements that induce pain. The BioVid heat pain dataset is an example of an experimental pain dataset, containing images of participants experiencing pain due to heat stimuli.

UNBC-McMaster Shoulder Pain Dataset

The dataset, previously described in Section 4.3.2, encompasses image sequences of 25 individuals experiencing shoulder pain while performing various arm movements. Each image is annotated with a PSPI score ranging from 0 to 15. As mentioned earlier, these images originate from videos and therefore exhibit strong similarities between consecutive frames. To capture the diverse range of pain intensities, a down-sampling method inspired by Zhao et al. [2016] and Xiang et al. [2022] was employed to eliminate redundant images. The down-sampling criteria were as follows:

(a) Each participant must have at least one image expressing pain and one image without pain.

(b) For sequences with the same pain intensity for five consecutive frames, only the first frame is retained.

Same as before, the images in the down-sampled dataset belonging to four participants (who consented to image publication) were assigned to the test set. Images from one randomly selected participant were reserved for validation, and the remaining 20 participants' images formed the training set.

BioVid Heat Pain Dataset

This analysis utilizes Part A of the BioVid dataset [Walter et al., 2013], which contains short videos capturing the facial expressions of 87 participants undergoing heat pain stimuli. Each participant is subject to five conditions (no-pain and four increasing pain intensities) and contributes 20 short videos (5.5 seconds in length) for each of the five conditions. Research findings indicate that the initial two pain intensities often fail to elicit a noticeable facial response [Werner et al., 2017]. Therefore, for pain detection, only videos labeled as baseline (no-pain) and the highest pain intensity were considered. Additionally, Werner et al. observed that facial activity associated with the highest pain intensity intensifies around the 4-second mark. Accordingly, the frame captured at the 4-second mark in each video was selected as a representative image. To eliminate participants with minimal or

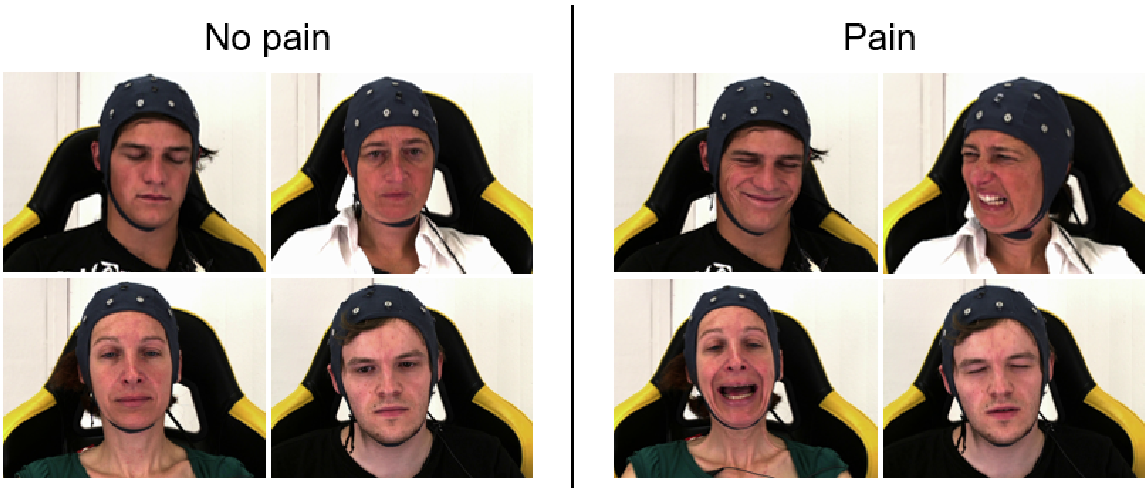


Figure 4.14: Some examples of pain and no-pain images from the BioVid heat pain dataset. The copyright remains with the dataset creators [Walter et al., 2013].

no facial response to the stimuli, the recommendations from the BioVid dataset creators [Walter and Al-Hamadi, 2022] were followed, leading to the exclusion of 20 participants. This results in images of 67 participants, from which 15 are designated for testing, 5 for validation, and the remaining images formed the training set. Example images from the pain and no-pain classes are presented in Figure 4.14.

4.5.2 Clinical and Experimental Pain Models

After selecting representative images from the videos or image sequences, the resulting datasets are relatively small with around 1000 to 2000 images. To address the challenge of limited dataset size, the models were trained using the transfer learning approach described in Section 4.3.2. Both the clinical and experimental pain models employed the AffectNet emotion recognition model as the source model for transfer learning. Both models functioned as binary classifiers, capable of distinguishing between pain and no-pain images. Both models underwent full fine-tuning, meaning that none of the layers were frozen to preserve the emotion recognition weights.

Pre-processing was applied to all input images, involving detection and extraction of the facial region using OpenCV. Subsequently, the images were resized to conform to the VGG16 input size of 224 × 224 pixels. The same data augmentation techniques employed in Section 4.3.2 – rotation, height shift, width shift, shear, zoom, and horizontal flip – were applied to the training images of both datasets. Both models utilized the SGD optimizer (learning rate = 0.01) and focal loss function ($\gamma = 2$), consistent with the previous training methodology. The samples selected from the BioVid heat pain dataset were evenly distributed between pain and no-pain images. In contrast, the down-sampled UNBC dataset exhibited an imbalance, with the number of pain images considerably exceeding no-pain images. Therefore, a weighted focal loss function was employed for the clinical pain model during the training phase. The weighting scheme proposed by Cui et al. [2019] was implemented, with the hyperparameter β empirically set to 0.99.

The clinical pain model trained on the UNBC shoulder pain dataset achieves an ac-

Paper	Accuracy	F1-score
Werner et al. [2016]	0.72	-
Yang et al. [2016]	0.60	-
Kächele et al. [2017]	0.66	-
Zhi and Wan [2019]	0.62	-
Othman et al. [2019]	0.66	-
Thiam et al. [2020]	0.69	-
Gkikas and Tsiknakis [2023b]	0.73	-
This chapter	0.70	0.69

Table 4.8: Performance (accuracy and F1-score) of BioVid heat pain detection models from the literature

curacy of 83%, which falls below the previously reported 91-93% accuracy range. This difference highlights the impact of image selection and class weighting on a model's performance. Further inspection of pain detection performances from Table 4.7 revealed that models trained on balanced datasets tend to achieve higher accuracy. Aspects such as image selection, pre-processing, and class weighting schemes are crucial for improving the performance of pain detection models. However, since the goal of the experiment is to compare the models trained on different types of pain datasets, the clinical pain model was trained using the image selection method that aligns with existing literature. The accuracy achieved by the clinical pain model described in this section remains comparable to other studies addressing similar classification tasks with an imbalanced training set.

The experimental pain model trained on the BioVid heat pain dataset achieved an accuracy of 70%, and F1-score of 69%. This performance aligns with the accuracy reported in other studies that utilize face images from BioVid dataset for pain detection, as shown in Table 4.8. This table compares the experimental pain model's performance to other similar models, all of which were trained as binary classifiers to detect presence of pain in BioVid facial images and evaluated on unseen participants. The works were identified through Scopus[4] database search.

The clinical pain and experimental pain models developed in this study serve as representative examples of pain detection models trained on clinical and experimental pain datasets. This enables the systematic comparison of their learned representations, shedding light on whether humans manifest pain differently in clinical and experimental settings.

4.5.3 Cross-Dataset Evaluation

Many studies have introduced automatic pain detection models that exhibit high performance within the context of their training datasets. However, cross-dataset evaluations

[4] `https://www.scopus.com/`, Query: TITLE-ABS-KEY (pain AND (detect* OR recogni* OR predict* OR estimat* OR classif* OR learn*) AND biovid)

have been less thoroughly investigated. According to Othman et al. [2019], one potential reason for this disparity is that even well-trained models may not perform well when deployed on another dataset. To comprehensively assess the generalizability of the pain detection models presented in this chapter, cross-dataset evaluations were conducted alongside the standard within-dataset evaluations.

The within-dataset evaluations, as described in Section 4.5.2, involved calculating recall, F1-score, and accuracy for both models using dedicated test sets from their respective datasets. These evaluations were conducted within the context of the datasets the models were trained on, evaluating their ability to detect pain expressions from unseen participants within the same pain stimuli and recording conditions. To address the limitations of within-dataset evaluations, cross-dataset evaluations were conducted using the test set from the BioVid heat pain dataset for the clinical pain model and the test set from the UNBC shoulder pain dataset for the experimental pain model. These cross-dataset evaluations assess the models' ability to detect pain caused by different pain stimuli and under different recording conditions. As mentioned previously, the potential pain stimuli in an industrial context can be clinical or experimental type. So, it is important to perform the cross-dataset evaluation using datasets representative of these types of pain to ensure that the model is robust in not only detecting pain in unseen participants but also in different pain stimuli and recording conditions.

The outcomes of within-dataset and cross-dataset evaluations for the clinical and experimental pain models are summarized in Table 4.9. The results of the cross-dataset evaluations revealed that the clinical pain model exhibited a considerable decline in performance when evaluated on the BioVid heat pain dataset, demonstrating its limited generalizability to different pain stimuli. Specifically, the low recall for the no-pain class is the primary reason for the decline in performance. In contrast, the experimental pain model maintained consistent performance across both datasets, suggesting that it has learned more generic pain features during training. This consistent performance suggests the model's potential for wider applicability in industrial settings where pain detection may involve varying pain types and recording conditions.

4.5.4 Visual Analysis

While the cross-dataset evaluation suggests differences in the learned feature representations of the two models, it is unclear whether these differences arise from variations in human pain responses to different pain types or if the clinical pain model has acquired dataset-specific features. To address this gap, this section employs the approach outlined in Section 4.4 to evaluate the learned features and gain insights into the pain expressions in the two datasets. The first step involved generating saliency maps to facilitate visual comparison of learned feature representations. Saliency maps were generated for each model to highlight the areas of the input image that each model considers to be relevant for activating the pain prediction neuron.

The test set images from both the UNBC and BioVid datasets were utilized to generate saliency maps. For each image, two saliency maps were created – one from the clinical pain model and the other from the experimental pain model. Since the objective was to

Clinical (UNBC) model							
Test Images	Recall			F1-score			Accuracy
	No-pain	Pain	Avg.	No-pain	Pain	Avg.	
UNBC (within)	0.71	0.88	0.80	0.72	0.88	0.80	0.83
BioVid (cross)	0.28	0.90	0.59	0.41	0.69	0.55	0.59
Experimental (BioVid) model							
Test Images	Recall			F1-score			Accuracy
	No-pain	Pain	Avg.	No-pain	Pain	Avg.	
BioVid (within)	0.87	0.54	0.70	0.74	0.64	0.69	0.70
UNBC (cross)	0.80	0.61	0.71	0.60	0.72	0.66	0.67

Table 4.9: Within-dataset and cross-dataset performances of clinical (UNBC) and experimental (BioVid) pain models.

identify disparities in the learned representations of the two models, the raw saliency maps were subtracted from each other to highlight differences. To enhance human visibility, the subtracted saliency maps were normalized between 0 and 1. This process yielded two sets of saliency maps – one highlighting the areas that the experimental pain model focused on more than the clinical pain model (‘Exp > Cli’ in Figure 4.15) and the other highlighting the areas more relevant for the clinical pain model than the experimental pain model (‘Cli > Exp’).

Figure 4.15 presents some of the test images and their corresponding saliency map differences. Upon inspection, it was observed that the clinical pain model tends to focus on the eye region, particularly on closing of the eyes. In contrast, the experimental pain model demonstrates a stronger focus on the mouth area, particularly on visibility of teeth. This observation leads to the hypothesis that the clinical pain model may be biased towards recognizing closed eyes as a pain indicator, while the experimental pain model may be biased towards detecting visible teeth as a pain indicator.

4.5.5 Embedded Concept Detection

The visual inspection of saliency maps indicated two distinct concepts (eye closure and visibility of teeth) that were differently emphasized in the learned feature representations of the clinical and experimental pain models. To validate whether the models’ learned representations exhibited significant disparities regarding these identified concepts, an embedded concept detection technique similar to Section 4.4.3 was employed. For each concept, a binary linear classifier was trained using the pain detection models. The model network acted as a feature extractor, relying on the output of the last pooling layer to generate feature vectors for each input image. The feature vectors extracted using the clinical and experimental pain models represent the latent representations of these models. The embedded concept detection approach assesses whether the models’ learned representations exhibited statistically significant differences in capturing the identified concepts.

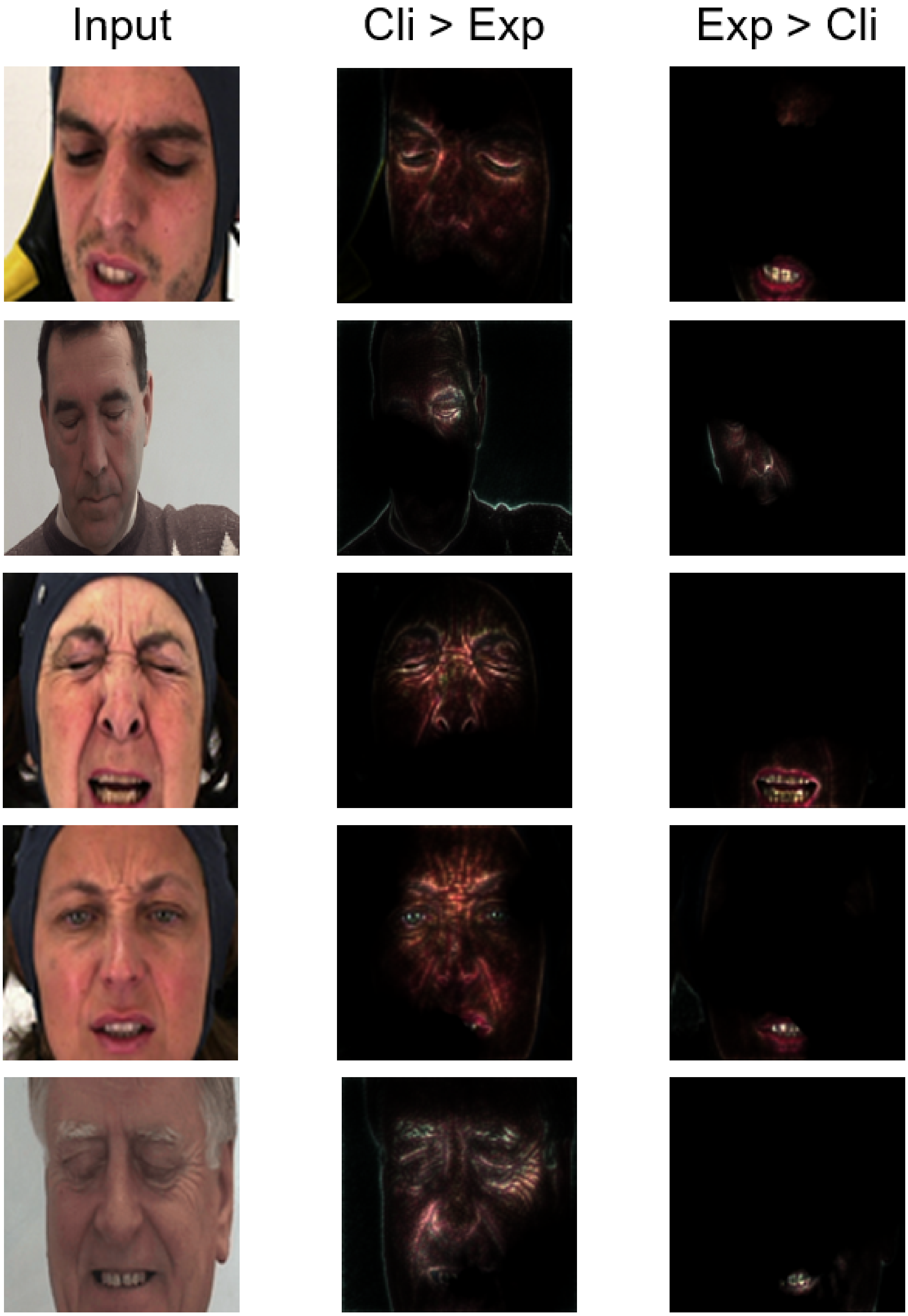


Figure 4.15: Subtracted saliency maps highlighting the differences in learned representations of experimental pain (BioVid) and clinical pain (UNBC) models. 'Cli' is short for Clinical pain model and 'Exp' is short for Experimental pain model. The copyright remains with the authors [Prajod et al., 2022a].

The test images from the UNBC and BioVid datasets were utilized to train the binary classifiers for the identified concepts. These images were manually annotated for the presence (0 or 1) of the identified concepts by two annotators. To ensure the reliability of annotations, two annotators independently assessed each image. In instances where the two annotators disagreed on the presence label, the images were presented to a third annotator for further evaluation. The final label was determined through a majority vote of the three annotators.

To assess the models' ability to capture the concept of closed eyes, the images were divided into two categories: those where participants had both eyes closed and those where they did not. Following the methodology outlined in Section 4.4.3, two binary classification SVMs were trained to identify closed eyes in face images. One SVM utilized feature vectors derived from the clinical pain model, while the other employed feature vectors from the experimental pain model. The classifiers were trained using 2-fold cross-validation, and the average F1-score across the two folds was calculated as a performance measure. This training process was repeated for 500 iterations using different random seeds for fold image selection and weight initialization. Finally, a significance test (paired t-test) was conducted to compare the 500 averaged F1 scores of the two SVMs. The SVMs trained using clinical pain model features achieved an average F1-score of 81.6%, while SVMs trained using experimental pain model features yielded a lower average F1-score of 78.4%. The observed difference was statistically significant, with a p-value of less than 0.001, suggesting a substantial distinction between the two pain detection models in their handling of the closed-eyes feature.

An identical procedure was implemented to evaluate the concept of visible teeth, which was also identified as potentially differentiating between the two pain detection models. The images were categorized into two groups based on whether the teeth were at least partially visible or not. In line with the approach employed for closed eyes, a pair of binary classification SVMs were trained using the pain detection models as feature extractors. The classifiers underwent 2-fold cross-validation over 500 iterations. The SVMs trained using clinical pain model features yielded an average F1-score of 73.5%, whereas SVMs trained using experimental pain model features achieved an average F1 score of 82.5%. The SVMs trained using experimental pain features were significantly better in distinguishing images with visible teeth compared to SVMs trained on clinical pain features (p-value: < 0.001). This result suggests that the visibility of teeth was more effectively embedded in the learned representations of the experimental pain model compared to the clinical pain model.

The outcomes of the two embedded concept detection support the initial hypothesis derived from saliency map analysis, confirming that the experimental pain model is more attuned to the concept of visible teeth, while the clinical pain model focuses more on closed eyes.

4.5.6 Insights

One interesting finding is that the clinical pain model exhibits poor performance in cross-dataset evaluation, despite performing well on the clinical pain dataset. In contrast, the experimental pain model demonstrates consistent performance across both datasets. A closer examination of Table 4.9 reveals that a significant contributor to the clinical pain model's decline in performance is the misclassification of no-pain images from the experimental pain dataset. As supported by the saliency maps and embedded concept detection results, the clinical pain model places greater emphasis on the eye region, particularly closed eyes. This observation aligns with the results reported by Kunz et al. [2019], who observed that closed eyes (AU 43) were more prevalent in clinical pain datasets compared to experimental pain datasets when analyzing both pain and no-pain images from a dataset.

This observation along with the findings of Werner et al. [2016] and Dai et al. [2019] provides a plausible explanation for the clinical pain model's misclassification of no-pain images. Werner et al. investigated various facial activity descriptors from the BioVid dataset to predict pain. They observed that eye closure is less pertinent in predicting pain compared to other features. This observation was attributed to the fact that some participants close their eyes even during no-pain videos. Analysis of the annotations generated for embedded concept detection revealed that approximately 20% of the no-pain images from the experimental pain test set were annotated as closed eyes. This may lead to experimental pain model relying less on closed eyes as a feature for detecting pain.

Dai et al. inspected videos of 20 random participants from the BioVid dataset and found that many of them closed their eyes for most of the experiment. In contrast, the participants from the UNBC dataset tend to look at the camera and usually close their eyes while in pain. This could lead to the experimental pain model strongly associating closing of eyes to pain expression. Hence, the experimental pain dataset (BioVid) containing no-pain images with closed eyes become challenging samples for the clinical pain model.

The inspection of saliency maps also revealed that the experimental pain model pays more attention to the mouth area, especially the visibility of teeth. This concept aligns with the pain pattern of 'open mouth', one of the four facial pain patterns identified by Kunz and Lautenbacher [2014]. They linked AUs 25, 26, and 27 to the open mouth pattern. However, as Werner et al. [2016] pointed out, these AUs are not included in the calculation of PSPI scores. The clinical pain dataset (UNBC dataset) is annotated based on PSPI scores, whereas the experimental pain dataset (BioVid dataset) is annotated based on the temperature applied. Therefore, it is possible that an image in the clinical pain dataset exhibiting an open mouth might be labeled as no-pain if the relevant PSPI AUs are absent. Furthermore, the annotations revealed that approximately 90% of the visible teeth images originated from the experimental pain dataset. Although the overall number of images with visible teeth is relatively low, the results indicate that this bias is reflected in the trained models.

4.6 Reflections and Remarks

Given the scarcity of pain datasets from industry-like settings, this chapter focuses on assessing the generalizability of pain detection models trained on existing datasets. While many studies evaluate generalizability through within-dataset testing on unseen participants, this chapter employs a more rigorous approach to investigate whether the models learn generic pain features rather than dataset-specific patterns.

First, transfer learning is employed to train pain detection models by leveraging the learned feature representations of an emotion recognition model. This training method mitigates the risk of overfitting due to the limited size of pain datasets.

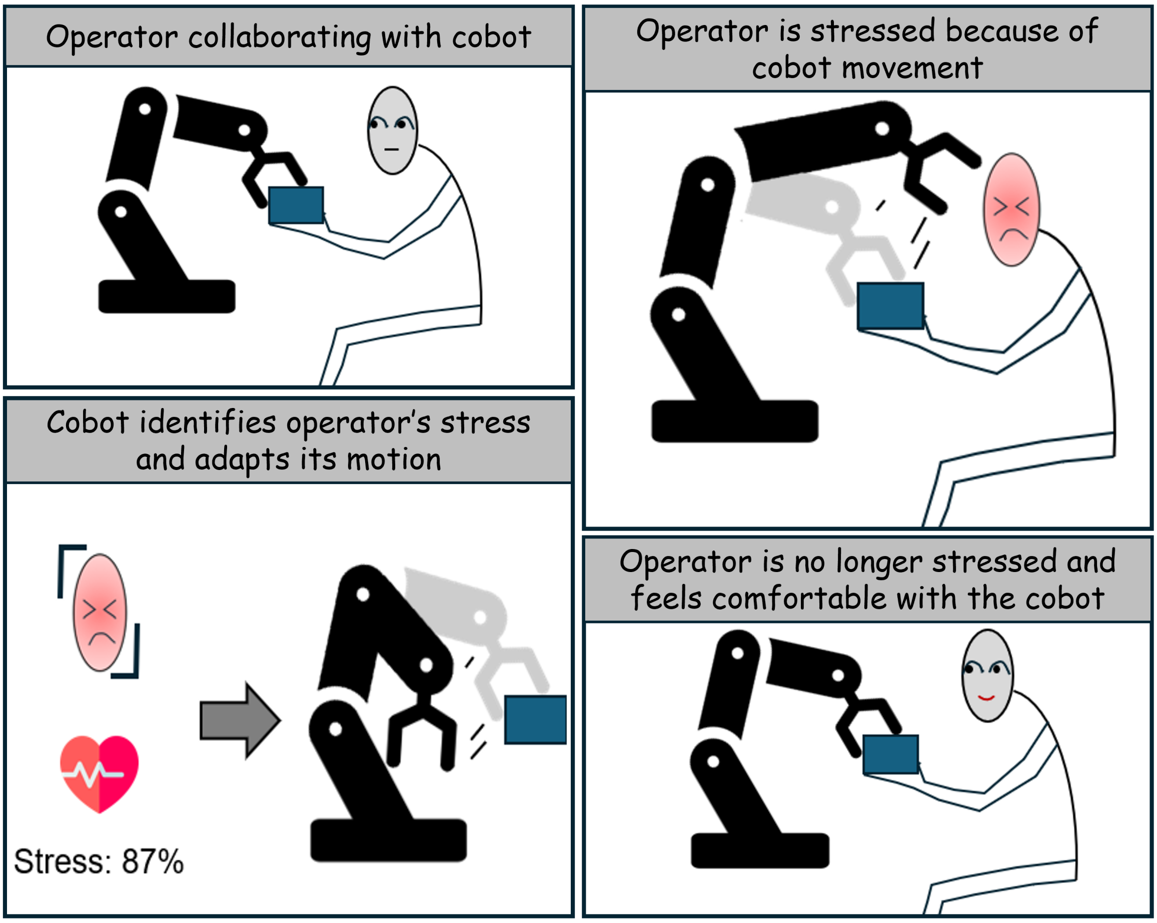
Second, an XAI-based approach is introduced to analyze the learned representations of the models. This approach involves generating saliency maps, manually inspecting these maps, and statistically verifying the hypotheses derived from manual inspection. This approach was used to investigate the impact of transfer learning from emotions to pain and to ensure that the fine-tuned pain model retains the learned representations of pain-related facial features.

Third, two pain detection models are trained on different pain datasets, and their cross-dataset performance is evaluated to determine which dataset leads to a more generalizable model. The XAI-based approach is employed here to identify dataset-specific features that contribute to cross-dataset performance drops.

The presented XAI-based approach is not limited to pain detection and can be applied to a wide range of machine-learning tasks. This versatility stems from the fact that the approach's steps are not tailored to pain detection, and can be applied to different domains.

Chapter 5

Stress Detection



Operator collaborating with cobot
Operator is stressed because of cobot movement
Cobot identifies operator's stress and adapts its motion
Stress: 87%
Operator is no longer stressed and feels comfortable with the cobot

Figure 5.1: A comic strip illustration of a hypothetical use case where stress detection during industrial Human-Robot Collaboration (HRC) improves the worker's well-being. The cobot moves very close to the operator during the HRC task. The cobot detects that the human operator is stressed due to its proximity. The cobot adapts its movement whenever it is close to the operator. This adaptation reduces the operator's stress and makes the collaboration experience more comfortable.

5.1 Overview

Stress detection and regulation in the workplace have numerous advantages, including improved productivity, job satisfaction, and worker well-being [Carr et al., 2011; Romero et al., 2016; Aidoo, 2016; Nicora et al., 2021; Mittal et al., 2022; Magtibay and Umapathy, 2023]. These advantages are particularly relevant in Industry 5.0 settings, where cobots work with human operators. In addition to typical workplace stressors like time pressure and social evaluation, specific aspects of the cobot's behavior, such as movement speed and proximity, can also induce stress in workers[Roy et al., 2020; Lu et al., 2022a]. For instance, in a hypothetical scenario illustrated in Figure 5.1, a cobot moving too quickly near a worker might trigger a stress response. Ideally, the cobot would be able to detect this stress and adjust its movements accordingly. Such adaptations not only reduce worker stress but can also increase trust and acceptance of the cobot [Simões et al., 2022; Lu et al., 2022a; Gervasi et al., 2023].

These potential benefits highlight the importance of developing automatic stress detection systems for Industry 5.0 HRC environments. However, a key challenge lies in the lack of publicly available stress datasets specifically collected in industrial HRC scenarios. As a result, researchers often rely on models trained on data from stressful non-industrial situations to detect stress in these settings. In such cases, it's crucial to assess the generalizability of these models and determine their applicability to broader real-world scenarios.

This chapter addresses the challenges of generalizability in stress detection models. It delves into various aspects of stress datasets, including the type of stressor, the sensors used for data collection, and the intensity of stress experienced. This research aims to identify which of these factors need to be aligned for developing stress detection models that have broader applicability to real-world scenarios. The chapter employs multiple models and evaluates their performance to draw broader insights that are not limited to a specific model. The experiments and results presented in this chapter have been previously published in the following papers:

- P. Prajod and E. André. On the generalizability of ECG-based stress detection models. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 549–554. IEEE, 2022

  [*I developed the machine learning models and performed the analyses. I also formulated research questions and derived insights.*]

- A. Heimerl, P. Prajod, S. Mertes, T. Baur, M. Kraus, A. Liu, H. Risack, N. Rohleder, E. André, and L. Becker. ForDigitStress: A multi-modal stress dataset employing a digital job interview scenario. *arXiv preprint arXiv:2303.07742*, 2023

  [*I contributed significantly to data curation, including data processing and feature engineering. I also trained the shallow machine learning models*]

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

[*I developed the machine learning models and performed the analyses. I formulated research questions and derived insights.*]

5.2 Previous Works

Recently, stress detection has gained traction in the field of affecting computing due to its prevalence in everyday situations like work, social interaction, etc. [Schmidt et al., 2019; Zamkah et al., 2020]. Moreover, detecting stress early is crucial due to its long-term health consequences.

5.2.1 Stress Datasets

Stress can be detected through psychological tools, behavioral patterns, and physiological signals [Giannakakis et al., 2019]. However, physiological signals are considered more reliable than the other methods due to their measurement issues (e.g., response biases, participants' control over their responses, etc.). Moreover, the growing popularity of unobtrusive wearable sensors further facilitates the acquisition of physiological signals.

This chapter focuses on heart-related signals - Electrocardiogram (ECG), Blood Volume Pulse (BVP), and Heart Rate Variability (HRV) - as they have been demonstrated to be reliable indicators of stress [Kim et al., 2018b; Gedam and Paul, 2020]. An overview of existing stress datasets that collected these signals is presented below. The datasets were identified through a Scopus database search[1].

Datasets collected in controlled laboratory settings often employ established stress-inducing tests such as the Stroop test [Stroop, 1935] and Trier Social Stress Test (TSST) [Kirschbaum et al., 1993]. The Stroop test elicits cognitive stress, and TSST induces social stress. Many studies collected physiological signals using Stroop test [Markova et al., 2019; Chen et al., 2021] or a combination of Stroop test and cognitive/arithmetic tests [Nakashima et al., 2016; Benchekroun et al., 2022; Xefteris et al., 2023]. While Markova et al. and Benchekroun et al. recorded both ECG and BVP signals, Chen et al., Nakashima et al., and Xefteris et al. acquired only one of those signals. All these studies, except Xefteris et al., collected Electrodermal Activity (EDA) signals as an indicator of stress response. Meanwhile, studies like Schmidt et al. [2018] and Sabour et al. [2021] employed the TSST technique to induce stress. Schmidt et al. collected multimodal signals using both chest-worn and wrist-worn devices, whereas Sabour et al. collected signals from a wrist-worn device.

A few studies [Parent et al., 2020; Coşkun et al., 2023] created datasets using video games to stress the participants. In these cases, a higher difficulty indicates a stressful condition. Although these studies employ a different stress elicitation method, they also induce cognitive stress.

The dataset by Iqbal et al. [2022] stands out for employing two widely different types of stressors (cognitive and social). They collected BVP signals utilizing both the Stroop test

[1] `https://www.scopus.com/`, Query: (TITLE (stress AND (dataset OR database)) AND TITLE-ABS-KEY (ecg OR electrocardio* OR bvp OR (blood AND volume AND pulse) OR ppg OR photoplethysmo* OR hrv OR (heart AND rate AND variability)))

and TSST to elicit stress. They observed that a majority of the participants had elevated heart rates in both situations. However, they did not perform any comparisons between the two responses.

Some researchers moved away from the established tests and simulated some real-world scenarios to induce stress. For example, Luig and Sontacchi [2014] collected ECG signals and speech data from pilots during an advanced flight simulation. Koldijk et al. [2014] designed a knowledge work scenario involving stressful conditions of interruptions and time pressure. They collected ECG and EDA signals in addition to behavioral data, such as computer logs and body pose. Campanella et al. [2024] simulated a LEGO assembly scenario and induced cognitive stress by incorporating some arithmetic tasks into the session.

Table 5.1 summarizes the datasets discussed above. Some datasets were collected in real-world scenarios without any explicit stressors. For example, Smets et al. [2018] collected ECG, EDA, and body temperature from office workers engaged in their daily routines. The dataset captured stress that naturally occurred during the day. Similarly, Jaiswal and Bara [2020] leveraged the final exam period of university students as the stressor. They collected physiological signals, including BVP, EDA, body temperature, and respiration. Since these studies did not utilize a controlled stressor, the type of stress experienced by the participants cannot be inferred.

5.2.2 Towards Generalizability of Stress Models

The datasets mentioned above led to the development of numerous uni-modal and multi-modal stress detection models [Can et al., 2019; Haque et al., 2024]. Some of these studies focused on comparing multiple models to determine the best model for stress detection. For example, Bobade and Vani [2020] compared the performances of stress recognition models such as Random Forest Classifier (RFC), Support Vector Machine (SVM), and Artificial Neural Network (ANN), trained on the hand-crafted multi-modal features from the WESAD [Schmidt et al., 2018] dataset. They showed that the simple ANN outperformed the other machine learning models. Similarly, Albaladejo-González et al. [2023] evaluated multiple HRV-based shallow models, including LDA (Linear Discriminant Analysis), RFC, SVM, and simple ANN. They trained their models on WESAD, SWELL-KW [Koldijk et al., 2014], and the combination of these datasets. They found that the ANN model performed consistently better than other models across all three datasets. In contrast, Gupta et al. [2023] found that the best model varied depending on the physiological signals. They investigated both uni-modal and multi-modal shallow models trained on the WESAD dataset. Although RFC was the best multi-modal stress classifier, LDA emerged as the best uni-modal model for ECG and BVP signals.

With the growing popularity of deep learning models, yet another question emerged: Do deep learning models trained on raw signals consistently outperform shallow models trained on hand-crafted features? To address this question, Hwang et al. [2018] and Ahmad et al. [2023] adopted a similar approach and designed ECG-based deep learning models to recognize stress. They demonstrated that these models yielded a better performance than HRV-based models on two different datasets. Similarly, Zhang et al. [2021] compared

	Paper	Stressor	Physiological Signals
ESTABLISHED TESTS	Nakashima et al. [2016]	Stroop test, Reading (Cognitive)	BVP, EDA
	Schmidt et al. [2018] (WESAD)	TSST (Social)	ECG, BVP, EDA, EMG, Resp, Temp
	Markova et al. [2019] (CLAS)	Stroop test (Cognitive)	ECG, BVP, EDA
	Parent et al. [2020] (PASS)	Video game (Cognitive)	ECG, EDA, Resp, Temp
	Chen et al. [2021] (MDPSD)	Stroop test (Cognitive)	BVP, EDA
	Iqbal et al. [2022] (Stress-Predict Dataset)	Stroop test, TSST, (Cognitive, Social)	BVP
	Sabour et al. [2021] (UBFC-Phys)	TSST (Social)	BVP, EDA
	Benchekroun et al. [2022] (MMSD)	Stroop test, Math (Cognitive)	ECG, BVP, EDA, EMG
	Coşkun et al. [2023] (AKTIVES)	Video game (Cognitive)	BVP, EDA, Temp
	Xefteris et al. [2023]	Stroop test, Math (Cognitive)	ECG, Resp
SIMULATED SCENARIOS	Luig and Sontacchi [2014]	Flight simulation (Cognitive)	ECG
	Koldijk et al. [2014] (SWELL-KW)	Knowledge work (Cognitive, Time)	ECG, EDA
	Campanella et al. [2024]	Assembly, Math (Cognitive)	BVP, EDA, Temp
	Heimerl et al. [2023]* (ForDigitStress)	**Job interview (Social)**	BVP, EDA

*Table 5.1: An overview of the existing controlled stress datasets that collected ECG or BVP signals. The entry marked with * is presented in this chapter.*

the performances of an ECG-based CNN-LSTM model with HRV-based machine learning models (SVM, RFC, etc.) using their own dataset. They noted that the deep learning model trained on 10-second segments of ECG data significantly outperformed the shallow models trained on HRV extracted from 60-second segments. Furthermore, Zhao et al. [2023] demonstrated that the superior performances of deep learning models extend to EDA and the combination of EDA and ECG signals.

All the models discussed until now, except Albaladejo-González et al. [2023], were trained and evaluated on the same datasets. Even the studies that utilized multiple datasets performed only within-dataset validations. There is a lack of evaluations to assess the generalization capabilities of stress detection models [Vos et al., 2023b].

Mishra et al. [2020] conducted an extensive cross-dataset study using four cognitive stress datasets - two had ECG signals, and the other two had BVP signals. In addition to arithmetic tasks, the ECG datasets had a startle response test and cold pressor task as stressors. Notably, the BVP datasets employed only arithmetic tasks as the stressor. The authors trained SVM models using HRV features extracted from the mental stress segments of these datasets and conducted cross-dataset evaluations. While the ECG-based HRV models performed well in detecting stress in each other's arithmetic tasks, they had a performance drop of 15 - 30% in predicting stress in the BVP datasets. The authors attributed this drop in performance, despite having the same stressor, to the difference in sensors. They also noticed an approximately 20 - 40% drop in the performance of ECG-based HRV models when detecting overall stress (including startle and cold pressor segments), even within the same datasets. Their findings suggest that the models trained on one type of stressor may not be efficient in detecting other stress responses.

As mentioned above, Albaladejo-González et al. [2023] trained HRV-based models on the WESAD and SWELL-KW datasets. While ANN outperformed other methods in within-dataset evaluations, they performed poorly in cross-dataset evaluations (WESAD models tested on SWELL-KW and vice versa). Moreover, the combining datasets did not improve the model's performance on individual datasets. However, it's important to note that they utilized a custom training and test data split, which may influence the results.

Benchekroun et al. [2023] trained two HRV-based models (RFC, logistic regression) on the MMSD [Benchekroun et al., 2022] and UWS [Velmovitsky et al., 2021] datasets. They tested the MMSD models using the UWS data and found that the F1-scores were 12 − 14% lower than the UWS models (from within-dataset evaluations). They further noted that the F1-score for stress class was very low (less than 50 %), meaning the models were not very efficient in detecting stressful instances.

Vos et al. [2023a] trained shallow models (RFC, SVM, XGBoost) using heart rate and EDA features from the SWELL-KW dataset and evaluated them on WESAD and NEURO [Birjandtalab et al., 2016] datasets. All three models showed poor cross-dataset performances. They also implemented an ensemble model (XGBoost + ANN) and repeated the cross-dataset evaluation. Although this ensemble model yielded a slight improvement, the performance was still poor (F1-score $<$ 50%). Furthermore, they trained the ensemble model using a combined dataset (SWELL-KW, NEURO, UBFC-Phys [Sabour et al., 2021]) and evaluated it on WESAD. While the accuracy increased slightly, the F1-score dropped further. Their experiments highlight the challenges of developing a generic stress model.

The above works primarily focused on assessing the generalizability of HRV-based models. However, Liapis et al. [2021] demonstrated that EDA-based shallow models also struggle with generalizability. They trained their models on the WESAD dataset and then evaluated them on their own dataset. Notably, their dataset contained subtle stress instances, unlike WESAD, implying that the stress intensity might further impact generalizability.

Baird et al. [2021] compared models trained on speech features from three social stress datasets. These datasets induced stress following the TSST technique. They predicted cortisol levels as a proxy for stress levels. In cross-dataset evaluations, the trends of predicted cortisol levels were aligned for the models, indicating compatibility between datasets. Due to the dataset compatibility, they suggested that training models using data from both these datasets could result in better-performing models.

5.2.3 Research Gap

The following three research gaps pertaining to generalizability of stress models have been identified through the literature review, and will be addressed in the subsequent sections.

- **Lack of generalizability assessments for ECG deep learning models**: As discussed in Section 5.2.2, ECG-based deep learning models often perform significantly better than HRV-based models. However, as evident from Table 5.2, studies that assess generalizability predominantly focus on HRV models. Hence, a gap exists in our knowledge of the generalization capabilities of deep learning models trained on ECG signals.

- **Lack of investigations into factors influencing generalizability**: As Table 5.2 highlights, most studies assess the generalizability of a model to determine if it is applicable in other stress scenarios. Some studies take a step further to evaluate whether combining datasets to train models improves stress detection performances. Although most studies observe low generalizability of stress models (see Section 5.2.2), few studies provide insights into plausible factors influencing the models' performance. For example, Mishra et al. [2020] highlighted the poor performance of mental stress models in detecting physical stress. Similarly, Liapis et al. [2021] noted the difference in stress intensities of the evaluated datasets. However, these studies did not further investigate these factors. Hence, a crucial question remains relatively unexplored - What factors or characteristics of the stress datasets need to match for cross-dataset applicability of models?

- **Need for a non-TSST social stress dataset**: Mishra et al. [2020] and Baird et al. [2021] are the only reviewed studies that reported good cross-dataset performances. While Mishra et al. observed this performance in arithmetic tasks, Baird et al. only evaluated social stress induced by the TSST technique. Since the stressor tasks involved in their chosen datasets are virtually the same, there is insufficient evidence to infer if matching the type of stressor would yield good cross-dataset performances. In other words - Would the models still yield good cross-dataset performance if the datasets

Paper	Input	Stressors	Aim
Mishra et al. [2020]	HRV features	Cognitive	Assess generalizability
Liapis et al. [2021]	EDA features	Social, UX stress	Assess generalizability
Baird et al. [2021]	Speech features	Social	Assess compatibility for combining datasets
Albaladejo-González et al. [2023]	HRV features	Social, Cognitive, Time	Assess generalizability, Combine datasets
Benchekroun et al. [2023]	HRV features	Cognitive, Daily routine	Assess generalizability
Vos et al. [2023a]	HR features, EDA features	Social, Cognitive, Time	Assess generalizability, Combine datasets
Prajod and André [2022]*	**Raw ECG**, HRV features	Social, Cognitive, Time	Assess generalizability, Combine datasets
Prajod et al. [2024b]*	HRV features	Social	Assess generalizability, Combine datasets, **Identify factors influencing generalizability**

*Table 5.2: An overview of the existing works that perform cross-dataset evaluations of their stress models. The entries marked with * are expanded in the subsequent sections of this chapter.*

had the same stressor types but induced by different methods? For answering this question using social stressors, the cross-dataset evaluations should involve a dataset that elicits social stress through methods other than TSST. As seen from Table 5.1, existing social stress datasets utilize TSST, which raises the need for a non-TSST social stress dataset.

5.3 Generalizability of ECG and HRV Models

Deep learning approaches are becoming increasingly popular in stress detection, with models trained on the ECG signals often performing better compared to shallow learning methods utilizing hand-crafted HRV features [Hwang et al., 2018; Zhang et al., 2021]. However, a critical question remains: can the deep-learning models perform equally well on other stress datasets?

This question is particularly relevant in domains like Industry 5.0, where publicly available stress datasets are scarce. If ECG-based deep learning models do not generalize well, their applicability in such scenarios might be limited.

This chapter addresses this question by investigating the generalizability of stress detection models. Five models are evaluated: two ECG-based deep learning models and three HRV-based shallow models.

5.3.1 Assessment Approach

The investigation follows a three-step approach:

1. **Within-dataset Assessment**: The first step assesses the stress detection performance of each model on its corresponding training dataset using leave-one-subject-out (LOSO) evaluation. This evaluation used two stress datasets to determine which models performed consistently well on data from unseen participants of the same dataset.

2. **Cross-dataset Assessment**: In the second step, the stress models trained on one dataset were evaluated using the other dataset and vice versa. This evaluation assesses to what extent these models can detect stress in new participants in different settings.

3. **Combining Datasets**: Finally, new models were trained on a combined dataset consisting of data from the two stress datasets, again using the LOSO technique. This step investigates potential improvements in models' performances due to the increase in both sample size and variations in the training data.

5.3.2 Datasets

This study leverages two publicly available stress datasets: WESAD [Schmidt et al., 2018] and SWELL-KW [Koldijk et al., 2014]. A brief comparison of the two datasets is presented in Table 5.3.

WESAD

The WESAD dataset is a multimodal stress and affect dataset containing various physiological signals, including ECG, EDA, and BVP. The data from 15 participants were collected using a chest-worn RespiBan and a wrist-worn Empatica E4 device. This investigation utilizes the ECG data recorded by the chest-worn device at 700 Hz.

The participants were subject to three conditions: neutral, amusement, and stress. In the stress condition, the participants experienced social stress induced by the TSST technique. The participants engaged in public speaking and mental arithmetic tasks while being evaluated by a three-member panel. To induce amusement, the participants watched selected funny video clips. The experimental sessions began with the neutral condition, followed by the stress and amusement conditions in alternating order. For each participant, the neutral condition lasted for approximately 20 minutes, the stress condition for 10 minutes, and the amusement condition for around 6.5 minutes.

This study focuses on stress detection, i.e., distinguishing between stress and no-stress samples. Following the labeling scheme proposed by the dataset creators, data from both neutral and amusement conditions were considered as no-stress samples.

SWELL-KW

The SWELL-KW dataset is also a multimodal stress dataset that contains two physiological signals, ECG and EDA. This dataset consists of data from 25 participants who engaged in typical knowledge tasks like writing reports and presentations. The ECG data was collected using the TMSI Mobi device at a sampling rate of 2048 Hz.

The participants underwent three experimental conditions: neutral, email interruptions, and time pressure. During the email interruption session, participants received eight emails, many irrelevant and some requiring responses. In the time pressure condition, participants had to complete the tasks within two-thirds of the allotted neutral session time. Like the WESAD dataset, the first session was always neutral, followed by the other two conditions in alternating order. The neutral and email interruption sessions lasted approximately 45 minutes, while the time pressure session was around 30 minutes long.

Notably, the participants did not report experiencing high stress in any of the three conditions. However, they indicated a higher temporal demand during the time pressure session. While training stress detection models, the dataset creators considered the data from email interruptions and time pressure sessions as stress samples and the neutral session as no-stress samples [Koldijk et al., 2016]. Therefore, this study follows the same labeling scheme for consistency. However, three participants were excluded due to missing data.

5.3.3 Data Processing

Down-sampling and Noise Removal

The WESAD and SWELL-KW datasets have different ECG signal sampling rates. HRV-based models can circumvent this difference through feature extraction. However, deep

	WESAD	SWELL-KW
Stressor	Social stress	Interruptions, Time pressure
ECG sensor	RespiBan, 700 Hz	TMSI Mobi, 2048 Hz
Avg. stress level	18.5/24 (STAI questionnaire)	3.5/10 (Likert scale)
Participants	15	22
Data duration (per participant)	stress: 10 mins, no-stress: 26.5 mins	stress: 75 mins, no-stress: 45 mins

Table 5.3: A comparison of some key characteristics of the WESAD and SWELL-KW stress datasets.

learning models utilize ECG signals and require fixed-length inputs. This restriction implies that both datasets should have ECG samples of the same length for cross-dataset validation. Hence, both datasets are down-sampled to 256 Hz for compatibility.

The ECG signals are affected by various sources of noise, including:

- **Baseline wander**: This is a low-frequency noise (0.5 - 0.6 Hz) caused by participants' body movements, respiration, etc., resulting in a drift in the ECG signal. A high-pass filter is typically applied to remove this noise [Kher, 2019; Limaye and Deshmukh, 2016; Luo and Johnston, 2010; Can et al., 2023].

- **Powerline interference**: This noise stems from the electromagnetic interference from the power supply of the sensor device. It is commonly removed by a band-stop or notch filter of 50 or 60 Hz, depending on the device [Kher, 2019; Limaye and Deshmukh, 2016; Luo and Johnston, 2010; Can et al., 2023].

- **EMG noise**: This is a high-frequency noise caused by muscle contractions and participant's movements. This noise is prominent in scenarios involving a lot of movement (e.g., exercise) and can be reduced using a moving average filter [Kher, 2019].

As described in Section 2.3.2 (refer to Figure 2.14), an ECG beat comprises P-wave, QRS complex, and T-wave. Stress detection primarily focuses on the QRS complex, which represents a heartbeat. Elgendi et al. [2010] suggest a frequency band of 8 - 20 Hz for optimal QRS signal-to-noise ratio. Therefore, a second-order Butterworth band-pass filter with this frequency range is applied. This filter removes most of the above-mentioned noises as their frequencies fall outside the chosen band.

Input Length

Previous studies [Hwang et al., 2018; Cho et al., 2019; Sarkar and Etemad, 2020; Zhang et al., 2021] have demonstrated that deep learning models trained on ultra short-term ECG signals perform well on stress detection tasks. Moreover, the two deep learning architectures considered in this study (see Section 5.3.5) were designed and validated using 10-second

ECG segments. Therefore, the deep learning models were trained using non-overlapping 10-second segments of filtered ECG data.

On the contrary, HRV feature extraction typically relies on longer ECG segments for reliable HRV feature calculation [Fang et al., 2022; Shaffer and Ginsberg, 2017; Schmidt et al., 2018; Pecchia et al., 2018; Pham et al., 2021]. This study utilizes 60-second segments with 50 seconds of overlap between consecutive segments. This overlap helps balance the number of training samples available for both ECG and HRV-based models.

Normalization

The ECG sensors used in the two datasets differ, potentially resulting in values recorded on different scales. In addition, the range of physiological recordings may vary from participant to participant [Braithwaite et al., 2013; Nkurikiyeyezu et al., 2019a; Sarkar and Etemad, 2020]. To mitigate the effect of these differences, participant-specific Min-Max normalization was applied to all inputs. It is important to note that normalization will not entirely remove the impact of different ECG sensors. For deep learning models, the filtered ECG data was normalized before using them as inputs to the models. On the other hand, participant-wise Min-Max normalization was applied to each HRV feature.

For real-time stress detection, the entire dataset wouldn't be available for normalization. Similar to Luong et al. [2020], this study used 5 minutes of neutral data to compute normalization parameters (minimum and maximum values) for each participant.

5.3.4 HRV Features

HRV features were extracted from the filtered ECG signals to train shallow models (see Section 5.3.6). First, the heartbeats were detected to derive the HRV signal, and then the corresponding features were computed.

The first step involved identifying peaks in the ECG signal, specifically the maximum amplitude within the QRS complex. This study utilized the algorithm proposed by [Elgendi et al., 2010] for R-peak detection. This algorithm is based on two key assumptions for healthy adults:

1. A QRS complex contains one and only one heartbeat

2. The duration of a typical QRS complex is in the range of 80 - 120 milliseconds

Once the R-peaks were identified, the time intervals between successive R-peaks were calculated to form the HRV signals. A total of 22 well-known features [Sriramprakash et al., 2017; Shaffer and Ginsberg, 2017; Schmidt et al., 2018; Pecchia et al., 2018; Giannakakis et al., 2019; Pham et al., 2021] were computed from the extracted HRV signals. These features belonged to the time domain (13 features), frequency domain (5 features), and poincaré plot characteristics (4 features). A brief description of each HRV feature is provided in Table 5.4. These features were calculated using the NeuroKit2 Python library [Makowski et al., 2021].

	Feature	Description
TIME	HR	Number of R peaks in 1 minute
	MeanNN	Mean of R-R intervals
	MedianNN	Median of R-R intervals
	MadNN	Median Absolute Deviation of R-R intervals
	StdNN	Standard deviation of R-R intervals
	CVNN	Ratio of StdNN to MeanNN
	IQRNN	Inter-Quartile Range of R-R intervals
	RMSSD	Root Mean Square of successive differences of R-R intervals
	StdSD	Standard deviation of successive differences of R-R intervals
	pNN50	% of successive differences of R-R intervals $> 50\ ms$
	pNN20	% of successive differences of R-R intervals $> 20\ ms$
	TINN	Triangular Interpolation of R-R intervals histogram
	HTI	HRV Triangular Index
FREQUENCY	LF	Power of low frequency band $(0.04\ Hz - 0.15\ Hz)$
	HF	Power of high frequency band $(0.15\ Hz - 0.4\ Hz)$
	LF/HF	Ratio of LF to HF power
	LFn	Normalized low frequency power, LF/total power
	HFn	Normalized high frequency power, HF/total power
POINCARÉ	SD1	Spread of HRV on Poincaré plot perpendicular to identity line
	SD2	Spread of HRV on Poincaré plot along identity line
	SD1/SD2	Ratio of SD1 to SD2
	S	Area of ellipse formed in the HRV Poincaré plot

Table 5.4: HRV features extracted from the ECG data and their descriptions.

5.3.5 ECG-based Models

This section details the two ECG-based stress detection models and their training parameters.

Deep ECGNet

This deep learning model was proposed by Hwang et al. [2018] for stress detection. It consisted of CNN (Convolutional Neural Network) layers for extracting features and LSTM (Long Short-Term Memory) layers for learning temporal patterns from the extracted features.

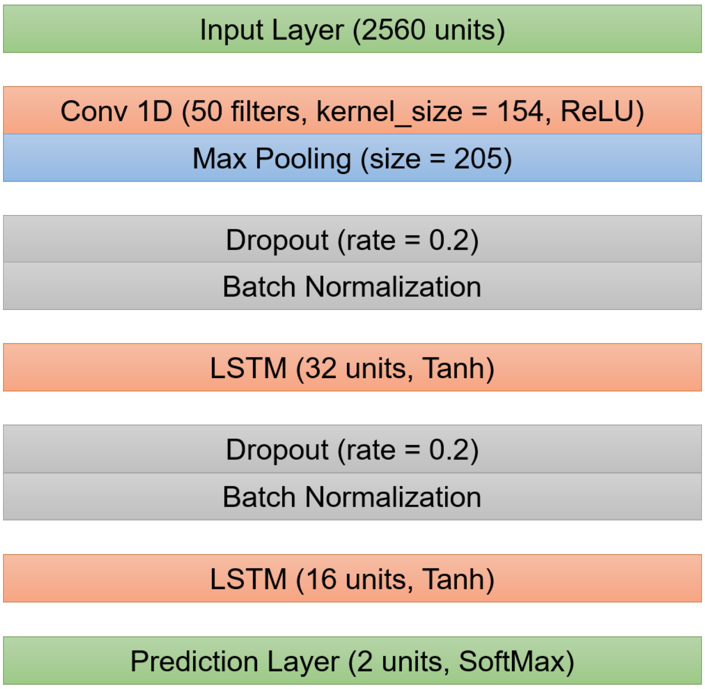


Figure 5.2: A visualization of the architecture of the Deep ECGNet Model

Figure 5.2 visualizes the architecture of the implemented DeepECGNet model. The model began with an input layer of size 2560 that accepts 10 seconds of ECG data sampled at 256 Hz. The input layer was followed by a convolutional block that consisted of a 1D convolutional layer, pooling layer, dropout layer (rate = 0.2), and batch normalization layer. The 1D convolution layer used the ReLU activation function and had 50 filters with a kernel size corresponding to 0.6 seconds of ECG data (kernel size = 154 for 256 Hz input). The pooling layer had a size of 205, which was equivalent to 0.8 seconds of data.

The output from the convolutional block was fed to a time series block with two LSTM layers: the first layer has 32 units, and the second layer has 16 units. A dropout layer (rate = 0.2) and a batch normalization layer were added between the LSTM layers. Finally, the model was connected to a prediction layer (Softmax activation).

The training process utilized the Adadelta optimizer with a learning rate of 1.0 and a weighted cross-entropy loss function to address class imbalance. The class-wise weights were determined based on the class frequencies of the training samples. The training lasted for 200 epochs with a batch size of 128 samples.

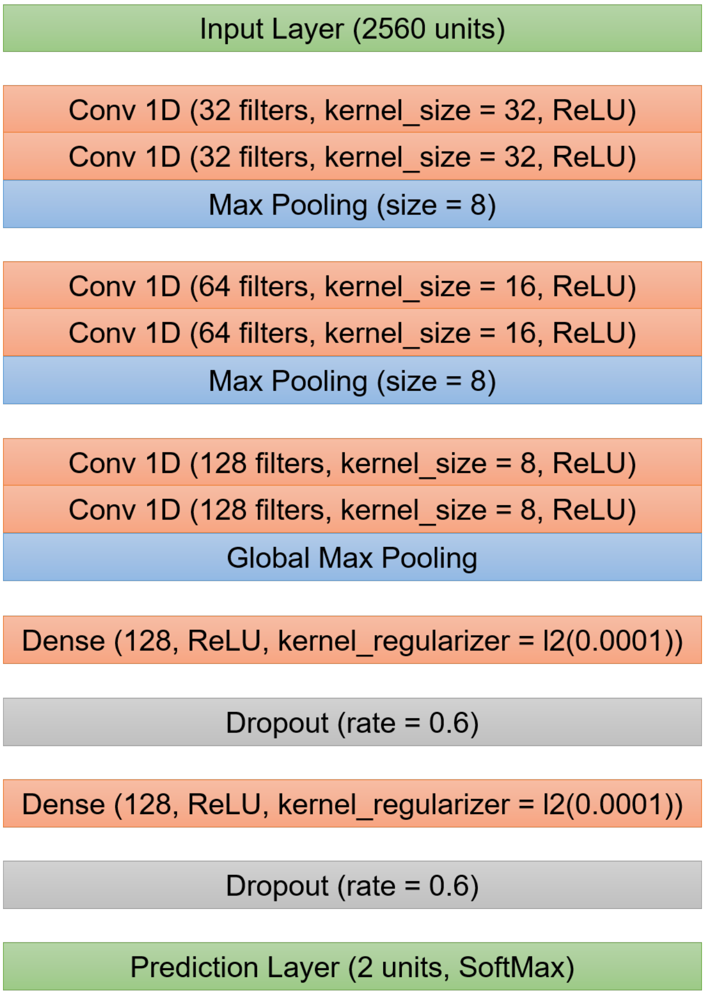

Input Layer (2560 units)
Conv 1D (32 filters, kernel_size = 32, ReLU)
Conv 1D (32 filters, kernel_size = 32, ReLU)
Max Pooling (size = 8)
Conv 1D (64 filters, kernel_size = 16, ReLU)
Conv 1D (64 filters, kernel_size = 16, ReLU)
Max Pooling (size = 8)
Conv 1D (128 filters, kernel_size = 8, ReLU)
Conv 1D (128 filters, kernel_size = 8, ReLU)
Global Max Pooling
Dense (128, ReLU, kernel_regularizer = l2(0.0001))
Dropout (rate = 0.6)
Dense (128, ReLU, kernel_regularizer = l2(0.0001))
Dropout (rate = 0.6)
Prediction Layer (2 units, SoftMax)

Figure 5.3: A visualization of the architecture of the ECG Emotion Model

ECG Emotion Model

This is a deep learning model proposed by Sarkar and Etemad [2020] for ECG-based emotion recognition on various datasets, including WESAD and SWELL-KW. Notably, unlike Deep ECGNet, this model relies solely on convolutional layers for feature extraction and classification. It does not incorporate recurrent layers (e.g., LSTM layers).

As depicted in Figure 5.3, the model consisted of three convolutional blocks, each containing two 1D convolutional layers and a pooling layer. The number of filters progressively increased from 32 to 64 to 128 across the blocks, while the corresponding kernel sizes decreased from 32 to 16 to 8. All convolution layers used the ReLU activation function. The pooling layers had a size of 8 with strides of 2. After the convolutional blocks, two dense layers were inserted, with 128 nodes in each. A dropout layer (rate = 0.6) followed each dense layer. The final layer of the model was the two-class prediction layer with Softmax activation.

Similar to Deep ECGNet, the model was trained using the Adadelta optimizer with a learning rate of 1.0, a weighted loss function, and training for 200 epochs with a batch size of 128.

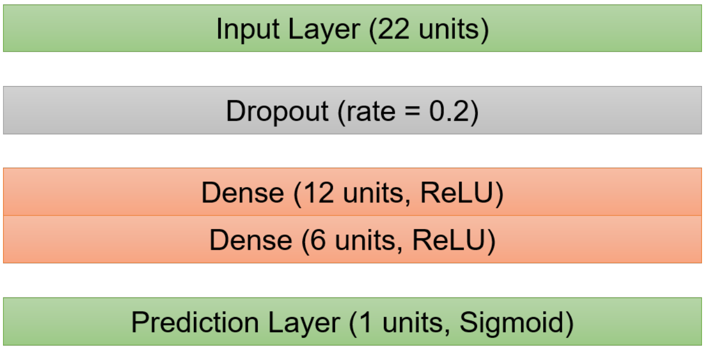


Figure 5.4: A visualization of the architecture of the Simple ANN Model

5.3.6 HRV-based Models

The following three shallow models were trained using the extracted HRV features (see Table 5.4).

RFC

This is an ensemble learning method that combines predictions from multiple decision trees for improved performance and reduced overfitting. Each tree is trained on a subset of the available training set. The final prediction is determined by aggregating the predictions from all the trees (e.g., majority vote). This strategy often results in a better performance, even if the individual decision trees are weak predictors.

Following the hyperparameters used in Schmidt et al. [2018], 100 decision trees (also called estimators) were trained, with a minimum of 20 samples for splitting a node within each tree. To account for the imbalanced sample distribution of the datasets, the “class_weight" hyperparameter was set based on the inverse sample frequencies.

SVM

This is a commonly used supervised learning method for binary classification tasks. During the training process of this model, the objective is to find a hyperplane within the feature space that separates the data points belonging to different classes.

Similar to previous research [Koldijk et al., 2016; Sriramprakash et al., 2017; Garg et al., 2021] on these datasets, an SVM with a Radial Basis Function kernel was employed. Like RFC, the “class_weight" hyperparameter was set inversely proportional to the sample frequencies.

Simple ANN

This is a simple feed-forward neural network (also called multi-layer perceptron), which has been growing in popularity for stress detection [Bobade and Vani, 2020; Zawad et al., 2023; Albaladejo-González et al., 2023].

As illustrated in Figure 5.4, the implementation followed an architecture consisting of an input layer, two hidden layers, and a prediction layer. The input layer received data represented as the normalized HRV features. A dropout layer (rate = 0.2) was included after the input layer to mitigate overfitting of the model. The two hidden layers (ReLU activation) followed the dropout layer, with 12 nodes for the first hidden layer and 6 nodes for the second hidden layer. This final layer outputs the classification result using a Sigmoid activation function.

Consistent with the other neural network models, this model was trained using the Adadelta optimizer (learning rate = 1.0) and weighted loss. It was also trained in batches of 128 samples for 200 epochs.

5.3.7 Assessment Results

All the models were trained and evaluated using the LOSO validation technique. In all three assessments, the performances of the models were measured in terms of accuracy and F1 scores.

Within-dataset Assessment

In this assessment, the models were trained and evaluated on the same datasets. Table 5.5 presents the LOSO evaluation results for the WESAD dataset, and Table 5.6 presents the results for the SWELL-KW dataset. These tables also include the performance of relevant existing models from the literature for comparison purposes. Although both datasets contain multimodal data, the comparison is limited to the models trained solely on ECG signals or HRV (derived from ECG) using LOSO validation. The binary classification models were identified through a literature search in the Scopus[2] database.

The deep learning models (Deep ECGNet and ECG Emotion Model) performed better than shallow machine learning models (RFC, SVM, Simple ANN) within their respective datasets. The margin of improvement is relatively moderate (within 5%) on the WESAD dataset. However, on the SWELL-KW dataset, the performance gap widens. Considering the F1-score and Accuracy, the Deep ECGNet emerges as the best model in both dataset categories.

Cross-dataset Assessment

Cross-dataset validations were employed to assess the generalizability of the models beyond their training datasets. In this assessment, models trained on the WESAD dataset were evaluated using the SWELL-KW data and vice versa. The cross-database performances of WESAD and SWELL-KW models are presented in the top and bottom parts of Table 5.7.

The results from the cross-dataset validation present a contrasting trend compared to the within-dataset evaluations. Deep learning models, which performed well in within-dataset evaluations, exhibit a significant drop in performance when tested on unseen data from the other dataset. In contrast, models trained on HRV features, particularly SVM for

[2] `https://www.scopus.com/`, Query WESAD: (TITLE-ABS-KEY ((detect* OR recogni* OR predict* OR classif* OR learn*) AND wesad AND (ecg OR electrocardiogram)) AND TITLE (stress)), Query SWELL: (TITLE-ABS-KEY ((detect* OR recogni* OR predict* OR classif* OR learn*) AND swell AND (ecg OR electrocardiogram)) AND TITLE (stress))

Model	F1-score	Accuracy
LDA [Schmidt et al., 2018]	0.813	0.854
LDA [Karan and Kaygun, 2021]	—	0.887
2D CNN (Spectrograms) [Liakopoulos et al., 2021]	0.794	0.824
Transformer, No Tuning [Behinaein et al., 2021]	0.697	0.804
Logistic Regression [Iqbal et al., 2021]	—	0.764
LDA [Gupta et al., 2023]	0.868	0.875
1D-CNN + 2D-ResNet [Ahmad et al., 2023]	0.875	0.877
RFC	0.813	0.863
SVM	0.832	0.871
Simple ANN	**0.859**	0.895
ECG Emotion Model	0.858	0.897
Deep ECGNet	0.857	**0.908**

Table 5.5: Results of LOSO evaluation of ECG or ECG-derived HRV models trained on the WESAD dataset. The first part tabulates the results of models from the literature and the second part shows the results of the five models considered in this chapter.

Model	F1-score	Accuracy
SVM [Koldijk et al., 2016]	-	0.589
Transformer, No Tuning [Behinaein et al., 2021]	0.588	0.581
RFC	0.644	0.670
SVM	0.609	0.639
Simple ANN	0.668	0.689
ECG Emotion Model	0.627	0.709
Deep ECGNet	**0.688**	**0.755**

Table 5.6: Results of LOSO evaluation of ECG or ECG-derived HRV models trained on the SWELL-KW dataset. The top part tabulates the results of models from the literature and the bottom part shows the results of the five models considered in this chapter.

Model	F1-score	Accuracy
Testing on SWELL-KW		
WESAD RFC	0.467	0.483
WESAD SVM	**0.535**	**0.538**
WESAD Simple ANN	0.478	0.49
WESAD ECG Emotion Model	0.395	0.411
WESAD Deep ECGNet	0.391	0.418
Testing on WESAD		
SWELL-KW RFC	**0.581**	0.637
SWELL-KW SVM	0.509	**0.647**
SWELL-KW Simple ANN	0.49	0.621
SWELL-KW ECG Emotion Model	0.342	0.385
SWELL-KW Deep ECGNet	0.392	0.415

Table 5.7: Results of cross-dataset evaluations of models from this chapter. The top part shows the performances of the WESAD models on the SWELL-KW dataset. The bottom part tabulates the results for the SWELL-KW models using the WESAD data.

WESAD and RFC for SWELL-KW, achieve the best overall performance (considering both F1 score and accuracy) during cross-dataset evaluations.

Combined Datasets

Finally, this study investigated whether combining the stress datasets (WESAD and SWELL-KW) could improve stress detection performances. Merging these datasets resulted in ECG data from 37 participants. The five models were trained and evaluated using LOSO validation on this combined dataset. The results are presented in Table 5.8.

Interestingly, combining the datasets did not lead to improved performance compared to training on individual datasets (see Tables 5.5 and 5.6). In fact, for the WESAD models, combining datasets resulted in a considerable decrease in both F1 scores and accuracies. The SWELL-KW models also have decreased performances, although it wasn't as substantial as the drop observed for WESAD models.

5.3.8 Insights

In within-dataset LOSO evaluations, the ECG models performed better than the HRV models on both datasets. Interestingly, the study by Zhang et al. [2021] reported similar findings on a self-collected stress dataset, where an ECG-based CNN-LSTM model outperformed HRV-based XGBoost model. Among the HRV-based models, the simple ANN achieved the best results within each dataset. This observation aligns with the findings of Bobade and Vani [2020] and Albaladejo-González et al. [2023], where a simple feed-forward network

Model	F1-score	Accuracy
Testing on WESAD		
RFC	0.758	0.793
SVM	0.732	0.796
Simple ANN	**0.758**	**0.813**
ECG Emotion Model	0.609	0.677
Deep ECGNet	0.692	0.711
Testing on SWELL-KW		
RFC	0.647	0.671
SVM	0.605	0.633
Simple ANN	0.657	0.677
ECG Emotion Model	0.593	0.683
Deep ECGNet	**0.695**	**0.739**
Testing on Combined		
RFC	0.692	0.720
SVM	0.657	0.699
Simple ANN	**0.698**	**0.732**
ECG Emotion Model	0.599	0.681
Deep ECGNet	0.694	0.728

Table 5.8: Results of LOSO evaluations of models trained on the combined dataset. The top part shows the performances of the models on the WESAD dataset, middle part displays the results for SWELL-KW dataset, and the bottom part tabulates the overall performances.

achieved better performance than other shallow machine learning methods (SVM, RFC) for multimodal stress detection.

The key takeaway from the cross-dataset validation is the better performance of HRV-based models compared to ECG-based models when tested on unseen data from a different dataset. This observation suggests that deep learning models might have learned dataset-specific features rather than generic stress patterns. The following three differences in the datasets (see Table 5.3) could potentially contribute to the drop in cross-dataset performances:

- **Stressor Differences**: The WESAD and SWELL-KW datasets utilize different stressors, leading to potential differences in the participants' stress responses.

- **Stress Levels**: The self-reported questionnaires indicate that the participants were quite stressed in the WESAD dataset, whereas the participants from the SWELL-KW dataset were not very stressed.

- **Sensor Differences**: The two datasets differ in the ECG sensors they employed (RespiBan vs. TMSI Mobi), potentially resulting in differences in signal characteristics. However, the HRV-based models use extracted features, which makes them less susceptible to this difference. This aspect could be the plausible reason for better cross-dataset performance of HRV models.

These factors could impact the generalizability of ECG-based deep learning models, which often require substantial training data that mimic the target scenario. Hence, focused studies with large datasets are needed to improve the robustness of these models. Based on these observations, this study recommends leveraging HRV-based models when the target application involves data that might differ from the training data. Deep learning models are suitable when the target scenario closely resembles the training dataset settings.

The experiment combining WESAD and SWELL-KW data revealed that simply merging datasets doesn't necessarily translate to improved stress detection performance. In fact, for the WESAD models, combining datasets resulted in a notable decrease in performance. This observation highlights the importance of data compatibility when considering such strategies.

5.4 Multimodal Social Stress - Dataset and Detection

Previous analyses revealed limitations in the generalizability of stress detection models, including those relying on HRV features. While HRV models demonstrated some robustness, factors contributing to better generalizability need further investigation. For this purpose, the ForDigitStress dataset was collected. Notably, this dataset employs a social stressor (similar to WESAD) to induce stress in participants. Moreover, the participants experienced high levels of stress, as evidenced by self-reported questionnaires and cortisol measurements (a biological marker of stress).

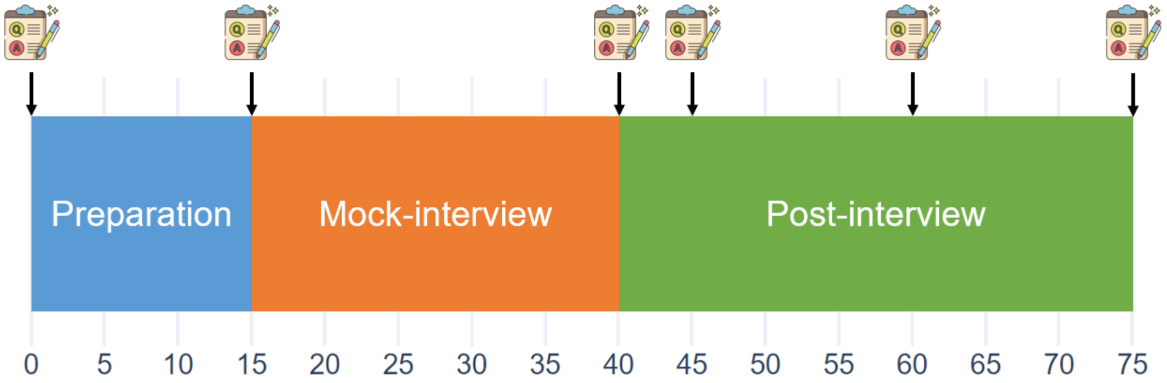


Figure 5.5: An illustration of the schedule for collecting stress questionnaires (and saliva samples).

5.4.1 Scenario and Protocol

The dataset collection setup employed a remote job interview scenario to induce stress and emotional arousal in participants. The setting replicated an online meeting by placing the participant and the interviewer in separate rooms, and they interacted via laptops. The participants played the role of job seekers interviewing for their dream positions. The participants were asked to submit their resumes in advance so that the interviewer could customize the questions depending on the participant. The interviewer questioned the participants on topics such as their strengths/weaknesses, salary expectations, and hypothetical job-related scenarios. Previous studies [Campisi et al., 2012; Gebhard et al., 2014; Becker et al., 2023] have shown that such simulated job interviews can elicit social stress.

The experiment began with a 15-minute preparation phase. During this time, participants were briefed about the data collection process. Then, the participants wore various sensors to measure their physiological responses throughout the experiment. While participants were preparing for the interview, the interviewer also used this time to finalize their interview questions. After the preparations, the mock interview began, which lasted for approximately 25 minutes. After the mock interview, a semi-structured qualitative interview took place. This follow-up interview aimed to discuss the participant's emotional experience during the mock interview. The follow-up interview lasted for 10 - 20 minutes, depending on the participant. The participants were encouraged to discuss specific moments that triggered feelings of stress.

Throughout the experiment, six saliva samples were collected from each participant. Two samples were collected before the mock interview began, and the remaining four samples were collected at intervals after the mock interview to capture the cortisol rise, peak, and eventual return to baseline. Figure 5.5 presents an illustration of the saliva sample collection schedule. At each saliva collection point, the participants also reported their perceived stress levels using a 10-point Likert scale ranging from "not stressed at all" to "totally stressed".

5.4.2 Data Acquisition

Sensors

The study employed a multi-sensor setup to capture the participants' responses to form the dataset. The following sensors were employed:

- Microsoft Kinect 2: This sensor recorded Full HD videos of the participant's upper body. The videos were recorded at 25 fps.
- Trust USB Headset: The participant wore a standard business USB headset to record their speech during the interview.
- IOM-biofeedback Sensor: This device collected BVP and EDA data of the participants at a sampling rate of 27 Hz.

Participants

The dataset included data from 40 healthy participants, with a gender distribution of 57.5% female, 40% male, and 2.5% diverse. Their age ranged from 18 to 31 years, with an average of 22.7 years (standard deviation of 3.2 years). The experiment yielded a substantial amount of multi-modal data. In total, 56 hours and 24 minutes of recordings were collected. The study was approved by Ethics Committee of Friedrich-Alexander-Universität Erlangen-Nürnberg (protocol no.: 21-408-S) and the data protection officer of the University of Augsburg.

5.4.3 Annotations

Ground Truth

Cortisol measurements and perceived stress levels were analyzed to assess the effectiveness of the scenario in inducing stress. Cortisol levels showed a significant change across the entire session. As visualized in Figure 5.6 (top), cortisol levels peaked five minutes after the interview and then gradually returned to baseline levels 35 minutes after the stressor (mock interview) ended. Perceived stress levels followed a similar pattern. Participants reported the highest stress levels immediately after the interview, with a subsequent decrease to baseline levels in the post-interview phase (as detailed in Figure 5.6, bottom).

Annotations

The cortisol and self-reported stress levels support the experiment's capability to induce stress. While the interview and post-interview phases can roughly be labeled as respective stress and no-stress samples, the participants were plausibly not stressed throughout the interview. In other words, the stress responses would be more pronounced during certain instances (e.g., questions) that were the stress triggers. Hence, a manual annotation process was employed for more nuanced stress labels.

The data was annotated by two experienced psychologists. The annotation process involved three steps. First, two annotators independently reviewed the videos using the NOVA tool [Baur et al., 2013], focusing on observable behaviors (video data) that might

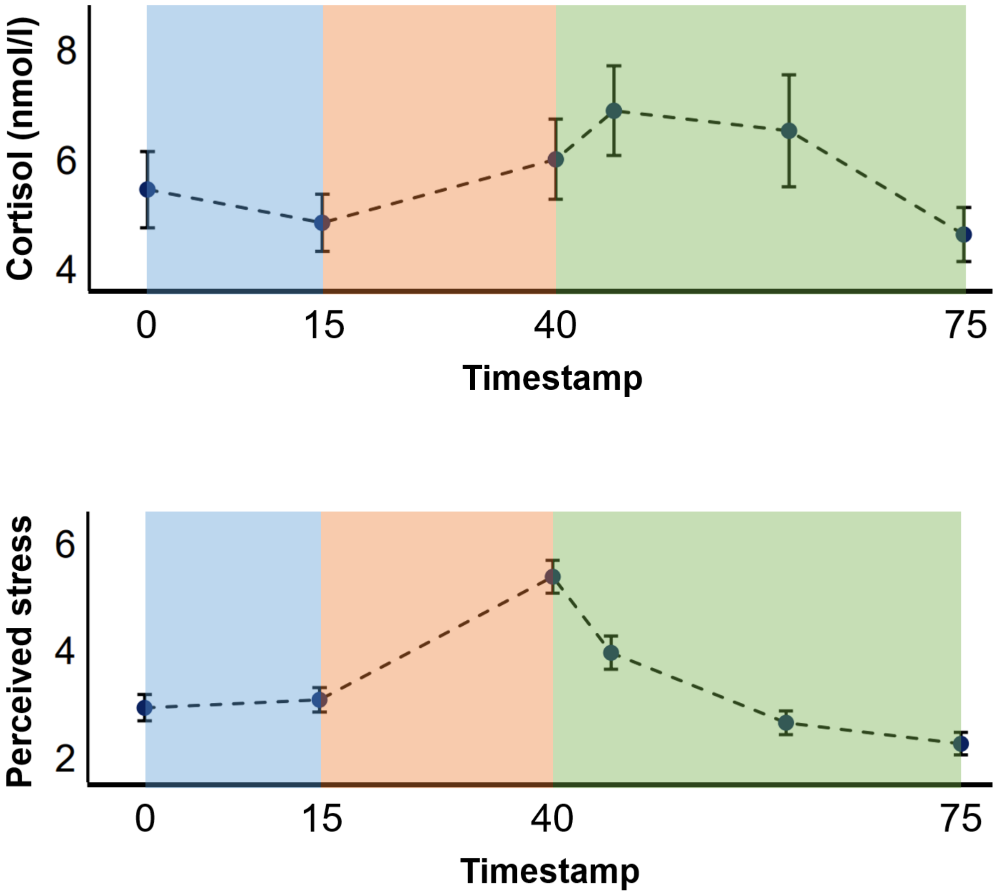


Figure 5.6: Plots showing the average cortisol level (top) and self-reported stress (bottom) throughout the experiment.

indicate stress or specific emotions mentioned by the participants (e.g., shame, anxiety, pride, etc.). Next, information from the follow-up interview, i.e., perceived stressful situations and emotions, was integrated with the behavioral observations. Based on this combined information, the annotators assigned discrete labels for stress and emotions to specific video frames. Finally, any disagreements in labels were discussed among the annotators and resolved to ensure consistency. Notably, there were no instances where the annotators couldn't assign time windows based on the participant reports, suggesting a strong correlation between subjective experiences and observable behaviors.

5.4.4 Data Processing and Feature Extraction

The dataset contains multimodal data that can be leveraged to detect stress. These data were first filtered to remove noise, and then relevant features were extracted. Features were extracted from BVP, EDA, video (face and body keypoints), and audio (speech) signals.

BVP

Features derived from HRV are often used in training stress detection models. During stress, heart rate increases and HRV decreases [Giannakakis et al., 2019; Pham et al., 2021]. Although ECG-based HRV features are commonly employed due to better signal quality [Umair et al., 2021], previous studies have demonstrated the effectiveness of BVP-based HRV features as an alternative [Namvari et al., 2022].

Like ECG, the BVP signal is susceptible to baseline wander and high-frequency noise. Hence, a band-pass filter (0.5 - 8 Hz) was applied to reduce these noises [Elgendi et al., 2013]. To derive the HRV signal from the BVP, the Systolic Peaks (see Figure 2.15) had to be detected. For this purpose, a peak-finding algorithm was employed to detect points that meet the following criteria:

(a) Amplitude threshold: The peaks had to be taller than a certain threshold. This threshold was set based on the distribution of peak heights in the entire signal.

(b) Distance between peaks: To avoid identifying every fluctuation as a peak, consecutive peaks had to be separated by a minimum interval of 0.333 seconds. This value corresponds to a maximum heart rate of 3 beats per second (180 beats per minute), the maximum heart rate observed by Kostis et al. [1982] during a physical stress scenario.

The same 22 features listed in Table 5.4 were computed using the BVP-based HRV. These features were calculated using BVP segments of 60 seconds. For consistency with other modalities like video that generate features for every frame (rate = 25 fps), the HRV features were calculated for every data point in the signal.

EDA

EDA is another commonly used physiological signal in stress detection [Healey and Picard, 2005; Schmidt et al., 2018; Koldijk et al., 2014; Nkurikiyeyezu et al., 2019b]. It indicates the activity of the sweat glands in the body, with higher EDA levels indicating increased stress. It can be separated into two components: Skin Conductance Level (SCL) and Skin Conductance Response (SCR) [Setz et al., 2009; Braithwaite et al., 2013; Schmidt et al., 2018; Giannakakis et al., 2019; Horvers et al., 2021]. SCL, or tonic component, is the slow-changing component indicating the underlying activity of the sweat glands. SCR, or phasic component, captures the rapid fluctuations in the EDA signal that occur in response to stress stimuli.

Before calculating the features, a 5 Hz low-pass filter was applied to the EDA signal to remove high-frequency noise [Setz et al., 2009; Schmidt et al., 2018; Horvers et al., 2021]. After noise removal, some statistical features (e.g., mean, standard deviation) of the filtered EDA signal were calculated [Schmidt et al., 2018; Nkurikiyeyezu et al., 2019b; Sriramprakash et al., 2017]. Next, the cvxEDA algorithm [Greco et al., 2015] was applied to decompose the filtered signal into its SCL and SCR components. Statistical features of the SCL and SCR components were computed. In addition, features associated with the SCR peaks, like number of peaks, duration, etc., were calculated [Healey and Picard, 2005; Nkurikiyeyezu et al., 2019b; Giannakakis et al., 2019].

Similar to BVP, these features were calculated for each data point in the signal, taking 60-second-long filtered EDA segments. Table 5.9 presents the 17 features that were extracted from the EDA signal.

Facial Action Units

Facial expressions are crucial for conveying emotions, making them the focus of automatic emotion recognition [Cohn and De la Torre, 2014; Dubey and Singh, 2016; Tarnowski et al., 2017]. These expressions are often represented in terms of facial action units. Recent research [Aigrain et al., 2015; Giannakakis et al., 2020, 2022] has shown promising results in using facial action units to predict stress levels. This study leveraged the Microsoft Kinect 2 camera to extract 17 facial action units (see Table 5.9) from the participant videos. These features were extracted for each video frame.

Body Keypoints

Previous research [Giakoumis et al., 2012; Aigrain et al., 2015; Chen et al., 2019] has demonstrated the effectiveness of analyzing body language and behavior in detecting stress. In this study, the OpenPose framework [Cao et al., 2017] was employed to extract features from the videos of participants. Although the Kinect camera provides body keypoints, it requires specific hardware to output 3D keypoints. This restriction may hinder comparison with other datasets that do not use this hardware. Hence, this study chose to utilize OpenPose as it can extract keypoints from videos recorded by any camera. However, unlike Kinect, OpenPose outputs 2D keypoints instead of 3D points.

A total of 24 features were extracted from 12 keypoints, as listed in Table 5.9. Since the participants were sitting, only the upper half of their bodies were visible throughout the session. Hence, only the keypoints corresponding to the upper body were considered.

Speech

Vocal cues are known to carry emotional information [Knapp et al., 1978]. Speech characteristics like pitch and speaking rate change depending on a person's emotions [Tao and Tan, 2005]. Stress, in particular, can trigger changes in the voice, such as a higher pitch (fundamental frequency) or increased vocal tremor [Mendoza and Carballo, 1999; Giddens et al., 2013; Lefter et al., 2015]. Several studies have successfully used acoustic features for automated stress detection [Lu et al., 2012; Kurniawan et al., 2013; Lefter et al., 2015; Han et al., 2018].

As a representation of participants' speech characteristics, this study extracted a set of well-established acoustic features called GeMAPS [Eyben et al., 2015]. A total of 58 features (see Table 5.9) were extracted from the recorded audio of the participants. These features included frequency and amplitude-related features (e.g., pitch, jitter, shimmer, loudness) and spectral features (e.g., Hammarberg Index, harmonic differences). The features were calculated for every one-second window of the speech data.

5.4.5 Training Baseline Models

Training Samples

As Section 5.4.3 described, the video recordings were analyzed, and stressful moments were annotated frame-by-frame. These frames formed the stress samples for training the model.

	Feature	Description
EDA	Signal	Mean, Standard deviation, Min, Max, Dynamic Range, Slope. Mean and Standard deviation of 1st derivative
	SCL	Mean and Standard deviation. Correlation with time
	SCR	Mean and Standard deviation. Number of peaks, Sums of peak amplitudes, peak durations, and area under the peaks
ACIAL ACTION UNITS	Jaw	Intensities of JawOpen, JawSlideRight
	Lips	Intensities of LipPucker, LipStretcherRight, LipStretcherLeft, LipCornerPullerLeft, LipCornerPullerRight, LipCornerDepressorLeft, LipCornerDepressorRight, LowerLipDepressorLeft, LowerLipDepressorRight
	Cheeks	Intensities of LeftCheekPuff, RightCheekPuff
	Eyes	Intensities of LeftEyeClosed, RightEyeClosed, RightEyebrowLowerer, LeftEyebrowLowerer
BODY	Face	x, y positions of nose, left eye, right eye, left ear, right ear
	Upper body	x, y positions of neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist
AUDIO	Frequency	Mean and Coefficient of variance of Pitch, Jitter, Formant 1, 2, 3 frequencies, Formant 1 bandwidth
	Amplitude	Mean and Coefficient of variance of Shimmer, Loudness, Harmonics-to-noise ratio
	Spectral	Mean and Coefficient of variance of Alpha Ratio, Hammarberg Index, Spectral Slope, Harmonic difference H1-H2, Harmonic difference H1-A3
	Temporal	Rate of loudness peaks, Pseudo syllable rate. Mean lengths and Standard deviations of voiced regions, unvoiced regions

Table 5.9: Features extracted from the EDA, video, and audio data. Features corresponding to facial action units and body keypoints were extracted separately from the video data.

Modality	Before PCA	After PCA
Action Units	17	10
EDA	17	9
BVP	22	10
Body Keypoints	24	8
Speech	58	19
Combined	138	49

Table 5.10: The length of feature vectors for each modality, before and after applying PCA.

Selecting appropriate no-stress data required some consideration. While both preparation and post-interview phases could be considered no-stress samples, speech primarily occurred during the follow-up interview. This interval began around 10 minutes after the mock interview and lasted at least 10 minutes (until the 20-minute mark). Therefore, data from this segment was used as the no-stress samples.

Since the stress annotations resulted in fewer samples than the no-stress segments, down-sampling was employed to address the class imbalance. It involved randomly removing no-stress samples to match the number of stress samples.

Dimensionality Reduction

The feature extraction process resulted in a substantial number of features from various modalities. A high number of features (also called dimensions of the feature vector) can pose challenges for some machine learning methods, especially when combining data from multiple modalities. To mitigate this issue, the Principal Component Analysis (PCA) technique was employed to reduce the dimensionality of the feature vectors. Reddy et al. [2020] showed that PCA effectively reduces feature dimensions without significantly impacting the performance of machine-learning models. PCA was applied to individual modalities and the combined data while retaining 95% of the variance in the data. The lengths of the resulting feature vectors were reduced for each modality, as described in Table 5.10.

For multi-modal stress detection, the following methods were used to combine features from all modalities:

- **Early PCA**: PCA was applied separately to each modality, and then the resulting feature vectors were combined. This method led to a combined feature vector with 56 dimensions (the sum of the reduced dimensions of each modality).

- **Late PCA**: The feature vectors from all the modalities were combined before applying PCA. This method resulted in a feature vector of 49 dimensions.

Like Reddy et al., all features were normalized using MinMax normalization before applying PCA.

Classifiers

Like the previous section (Section 5.3.6), three shallow models were trained. The hyperparameters of these models were chosen empirically for this dataset. All the models were trained and evaluated using the LOSO technique.

- **RFC**: This implementation used 100 estimators and a minimum sample size of 50 for splitting nodes.
- **SVM**: Due to the higher number of samples and feature dimensions, training an SVM with the radial basis kernel was computationally heavy. So, a linear SVM was trained instead.
- **Simple ANN**: Similar to the previous study, the ANN had an input layer, two hidden layers, and a final prediction layer. The size of the input layer can vary depending on the modalities used. Therefore, the number of nodes in the hidden layers depended on the input layer size. The first hidden layer had nodes equal to half the input size (rounded to the nearest multiple of 2). In turn, the second hidden layer had half the nodes as the first hidden layer.

  Moreover, the training process utilized the SGD optimizer (learning rate = 0.001) and the binary cross-entropy loss function. The ANNs were trained for a maximum of 100 epochs and used a batch size of 256 samples. The training utilized an early stopping mechanism (patience = 15 epochs) to prevent overfitting. This mechanism stopped training if the model's validation loss did not decrease for 15 consecutive epochs.

5.4.6 LOSO Evaluation Results

Table 5.11 presents the LOSO evaluation results of the three shallow models. The table shows the results for models trained on individual modalities and combined feature vectors (early and late PCA).

The results revealed that combining information from multiple modalities led to better stress detection performance compared to using individual modalities alone. However, early PCA yielded slightly better results across the tested models.

Notably, ANN emerged as the best shallow model in all individual modalities as well as in early and late PCA feature vectors. The ANN trained on early PCA features achieved the best stress detection performance (F1 score = 88.1%, accuracy = 88.3%).

When analyzing individual modalities, BVP (HRV features) consistently yielded the best performances across all models, followed by facial action units and body keypoints. Speech (GeMAPS) and EDA features showed the lowest effectiveness in stress detection, achieving F1 scores and accuracies 15-20% lower than HRV features.

5.4.7 Insights

Models trained on HRV features outperformed other individual modalities. This result aligns with previous research, which showed heart rate and HRV as reliable stress indi-

Modality	RFC	SVM	Simple ANN

Facial Action Units	F1 = 71.4,
	Acc. = 73.6
		F1 = 75.6,
		Acc. = 77.2
			F1 = 76.5,
			Acc. = 78.0

EDA	F1 = 54.2,
	Acc. = 57.1
		F1 = 57.6,
		Acc. = 58.9
			F1 = 60.2,
			Acc. = 61.3

BVP	F1 = 74.5,
	Acc. = 75.9
		F1 = 76.1,
		Acc. = 77.7
			F1 = 78.4,
			Acc. = 79.7

Body Keypoints	F1 = 59.4,
	Acc. = 63.6
		F1 = 69.8,
		Acc. = 73.4
			F1 = 76.4,
			Acc. = 79.5

Speech	F1 = 52.1,
	Acc. = 55.9
		F1 = 57.3,
		Acc. = 58.9
			F1 = 58.7,
			Acc. = 60.3

All modalities (early PCA)	F1 = 81.3,
	Acc. = 82.0
		F1 = 83.8,
		Acc. = 84.5
			F1 = 88.1,
			Acc. = 88.3

All modalities (late PCA)	F1 = 78.2,
	Acc. = 79.3
		F1 = 83.9,
		Acc. = 84.5
			F1 = 87.5,
			Acc. = 87.7

Table 5.11: LOSO evaluation results (F1-score and Accuracy) for classifiers on individual modalities and combined feature sets.

cators [Kim et al., 2018b; Gedam and Paul, 2020]. Notably, face and body pose features also achieved promising results. However, these features rely on the visibility of the face and body, which can be challenging in industrial settings [Mosberger and Andreasson, 2013; Tarabini et al., 2018]. For example, personal protective gear or the positioning of the worker might obstruct the camera's view.

EDA is another popular modality utilized in automatic stress detection. While some studies report high accuracies for TSST stress classification using EDA [Schmidt et al., 2018; Greco et al., 2021], the models trained solely on EDA features in this study achieved the second-lowest performance. A possible explanation is the time delay between stress triggers and the corresponding EDA response [Sjouwerman and Lonsdorf, 2019; Gradl, 2020; Callara et al., 2021]. The stress annotations in this dataset have a high temporal resolution, which may not account for the delayed EDA response.

Models trained on speech features (GeMAPS) yielded the lowest performance in this study. Similar research reported moderately better results [Kurniawan et al., 2013; Han et al., 2018], suggesting potential for performance improvement. However, it is worth noting that the applicability of speech-based stress detection is limited in industrial settings due to the noisy environments [Strazdas et al., 2020; Agati et al., 2020].

5.5 Generalizability of Social Stress Models

The WESAD dataset shares more similarities with the ForDigitStress dataset than the SWELL-KW dataset. In other words, WESAD and ForDigitStress are more compatible. Both datasets induced social stress, although the methods employed to elicit stress differed. The findings of Section 5.3 suggest two primary factors that affect the generalizability of HRV-based stress models: the type of stressor and intensity of stress. By comparing two datasets with the same stressor type, this study aims to assess the importance of matching stress types in a model's broader applicability.

5.5.1 Datasets

Modalities

This study leveraged two stress datasets, WESAD and ForDigitStress. As described in previous sections, both datasets induced social stress in participants. However, their stress elicitation technique differed: WESAD utilized public speaking and mental arithmetic, whereas ForDigitStress employed mock job interviews. While WESAD contains ECG and BVP signals for deriving HRV, the ForDigitStress dataset provides the BVP data. For consistency in modality across datasets, this study utilized the BVP signals from the datasets to extract HRV features.

Another reason for focusing on the same modality is to control the influence of sensor type. While ECG-based and BVP-based HRVs reflect similar physiological information related to heart activity, the models trained on these features might not perform equally well [Gupta et al., 2023]. This discrepancy can be attributed to the BVP signals being more prone to noise from body movements than ECG [Martinho et al., 2018]. This noise can affect the signal quality, which can, in turn, impact the stress detection performance.

The two datasets used different sensors to record the BVP data. The BVP signal in WESAD was collected using an Empatica E4 device with a sampling rate of 64 Hz. In the ForDigitStress dataset, BVP was recorded using an IOM-biofeedback sensor at a lower sampling rate of 27 Hz.

Samples

This study utilized all data from the WESAD dataset belonging to stress and no-stress (neutral and amusement) classes for training the stress detection models. While the ForDigitStress dataset has frame-by-frame stress labels, such labels are absent in WESAD. Aligning their stress labeling processes, the entire mock interview phase in ForDigitStress was considered stressful (excluding the first 5 minutes).

Further deviating from the original ForDigitStress labels, the post-interview segments from the 20-minute mark onwards were chosen as no-stress samples. This choice was primarily based on the observation that cortisol and perceived stress returned to baseline levels after the 20-minute mark. Moreover, the presence of speech was no longer a selection criterion. A summary of key characteristics of the datasets is presented in Table 5.12 for comparison.

	WESAD	ForDigitStress
Stressor	Social stress	Social stress
BVP sensor	Empatica E4, 64 Hz	IOM-biofeedback, 27 Hz
Avg. stress level	18.5/24 (STAI questionnaire)	5.4/10 (Likert scale) 6.5/10 (Cortisol)
Participants	15	40
Data duration (per participant)	stress: 10 mins, no-stress: 26.5 mins	stress: 20 mins, no-stress: 15 - 20 mins

Table 5.12: A comparison of some key characteristics of the WESAD and ForDigitStress datasets.

5.5.2 Training Models

Similar to the generalizability assessment in Section 5.3, three shallow models were trained on both datasets. The models utilized the same HRV features as before, listed in Table 5.4.

Data Processing

The same processing steps described in Section 5.3 for the ECG signal were also applied to the BVP signal. An exception was the down-sampling of both datasets to the same frequency. HRV features can be extracted from BVP signals sampled at different frequencies as long as the resolution is sufficient to identify individual heartbeats. So, the original sampling frequencies were retained.

Extracting HRV features from BVP signals followed the method described in Section 5.4.4.

1. **Noise Removal**: A band-pass filter (0.5 - 8 Hz) was applied to the BVP signals. This filter removed unwanted noise components, including baseline wander and high-frequency noise.

2. **Peak Detection**: A peak-finding algorithm was used to identify peaks in the filtered BVP signals. Peaks exceeding a certain amplitude and separated by a minimum time interval were identified as heartbeats.

3. **Input Length**: HRV features are calculated by analyzing 60-second segments of the filtered BVP data, with an overlap of 59 seconds. This sliding window technique generated one feature vector for each second of the recording.

The final step of data preprocessing involved normalizing the extracted HRV features using MinMax normalization. For each participant in the WESAD dataset, the normalization parameters were calculated using the initial 5 minutes of the neutral phase. These neutral segments represent a baseline state with minimal stress influence. In contrast, stress levels in the ForDigitStress dataset tend to decrease towards the end of the recording sessions. Therefore, the normalization parameters were calculated using the last 5 minutes of the participant's data.

Classifiers

Consistent with previous studies of this chapter, three shallow models were trained (RFC, SVM, and Simple ANN). The hyperparameters of these models were chosen empirically.

- **RFC**: This implementation used 200 estimators and a maximum depth of 5. Additionally, the class weights parameter was computed as inverse class frequencies.
- **SVM**: Due to the large sample size in this experiment, a linear kernel was utilized with the SVM. The tolerance parameter was set to 10^{-5}. The same class weights as RFC were used here.
- **Simple ANN**: Similar to Section 5.3.6, the ANNs had an input layer, dropout layer (rate = 0.2), two hidden layers (12 nodes and 6 nodes), and a final prediction layer. The models were trained for 200 epochs in batches of 256 using the SGD optimizer (learning rate = 0.001) and the weighted binary cross-entropy loss function. An early stopping mechanism (patience = 15 epochs) was also implemented.

5.5.3 Assessment Results

Like Section 5.3, the assessment involved within-dataset, cross-dataset, and combined dataset evaluations. The models were trained using the LOSO method, and their performances were measured in accuracy and F1-score.

Within-dataset Assessment

This assessment involved training and evaluating models on the same dataset. The assessment was repeated on two social stress datasets. The LOSO evaluation results for the WESAD dataset are presented in Table 5.13, along with the performances of existing BVP-based models evaluated using LOSO on the WESAD dataset. The models were identified through Scopus database search[3]. Table 5.14 presents the LOSO results for the ForDigit-Stress dataset.

The within-dataset LOSO evaluations revealed a consistent trend within each dataset: simple ANN achieved the highest performance, followed by RFC, and lastly, SVM. However, it's important to note that the performance differences between these models were relatively small, falling within a margin of 3%. Interestingly, the models also yielded similar performance levels on both datasets.

[3] `https://www.scopus.com/`, Query: (TITLE-ABS-KEY ((detect* OR recogni* OR predict* OR classif* OR learn*) AND wesad AND (bvp OR blood AND volume AND pulse OR ppg)) AND TITLE (stress))

Model	F1-score	Accuracy
LDA [Schmidt et al., 2018]	0.830	0.858
1D-CNN [Lisowska et al., 2021]	0.837	—
ANN (HRV) + CNN(raw) [Rashid et al., 2021]	0.862	0.886
1D-CNN [Zhang et al., 2023b]	0.528	0.570
LDA [Gupta et al., 2023]	0.805	0.82
RFC	0.768	0.826
SVM	0.763	0.792
Simple ANN	**0.780**	**0.829**

Table 5.13: Results of LOSO evaluation of BVP-derived HRV models trained on the WESAD dataset. The first part tabulates the results of models from the literature and the second part shows the results of the three models considered in this chapter.

Model	F1-score	Accuracy
Simple ANN (Table 5.11)	0.784	0.797
RFC	0.810	0.829
SVM	0.787	0.811
Simple ANN	**0.814**	**0.831**

Table 5.14: Results of LOSO evaluation of BVP-derived HRV models trained on the ForDigitStress dataset. The bottom part shows the results of the five models considered in this chapter.

Cross-dataset Assessment

The objective of cross-dataset assessment was to evaluate the generalizability of the social stress models. The LOSO models trained on one dataset (WESAD or ForDigitStress) were assessed on the other dataset. Table 5.15 shows the results of this cross-dataset evaluation. The top part of this table presents the performance of models trained on WESAD data when tested on ForDigitStress data. Meanwhile, the bottom part tabulates the performance of models trained on ForDigitStress data when tested on WESAD data.

The cross-dataset evaluation of social stress models revealed a significant improvement compared to the previous cross-dataset assessment in Section 5.3, which involved datasets with different stressors. Notably, there was no consistent best-performing model. SVM performed the best among the WESAD models, whereas RFC was the best among the ForDigitStress models. Despite being the best performer in within-dataset evaluations, simple ANN did not outperform the other models in cross-dataset evaluations. However, the performance differences between the models remained relatively small ($< 4\%$).

Model	F1-score	Accuracy
Testing on ForDigitStress		
WESAD RFC	0.731	0.740
WESAD SVM	**0.774**	**0.775**
WESAD Simple ANN	0.740	0.744
Testing on WESAD		
ForDigitStress RFC	**0.789**	**0.820**
ForDigitStress SVM	0.779	0.812
ForDigitStress Simple ANN	0.763	0.810

Table 5.15: Results of cross-dataset evaluations of models from this chapter. The top part shows the performances of the WESAD models on the ForDigitStressdataset. The bottom part tabulates the results for the ForDigitStress models using the WESAD data.

Combined Datasets

To explore the potential improvements due to a larger dataset, WESAD and ForDigitStress were merged, resulting in a substantial amount of data from 55 participants. Again, the LOSO method was utilized to train stress detection models on the combined dataset. The results are presented in Table 5.16. There is a slight increase in performance compared to models trained on individual datasets. The simple ANN model tested on WESAD data showed a notable improvement of approximately 3%.

5.5.4 Insights

As expected, the simple ANN achieved the best LOSO performance within each dataset (WESAD and ForDigitStress). Interestingly, all models trained on the ForDigitStress dataset outperformed the previously reported best BVP performance (Table 5.11). This improvement might be due to how no-stress samples were selected. For training baseline models with ForDigitStress (see Section 5.4.5), the no-stress samples were chosen based on the presence of speech in the post-interview phase. However, Figure 5.6 suggests that stress levels might not have entirely returned to baseline during this window, implying that some selected “no-stress" samples might still contain residual stress responses.

Unlike the SWELL-KW dataset used in Section 5.3, all models trained on both WESAD and ForDigitStress datasets achieved comparable LOSO results within each dataset. This observation suggests no inherent differences in the models’ stress prediction capabilities, a promising trend for cross-dataset evaluations.

The cross-dataset evaluation revealed good generalizability across datasets, meaning models trained on one dataset could be applied to the other without compromising performance. This result contrasts the findings in Section 5.3, where models struggled with generalizability. Two main dataset differences were identified that could have contributed to the poor generalizability of the HRV models: type of stressor and stress intensity. Both

Model	F1-score	Accuracy
Testing on WESAD		
RFC	0.783	0.829
SVM	0.765	0.821
Simple ANN	**0.808**	**0.857**
Testing on ForDigitStress		
RFC	0.812	0.831
SVM	0.780	0.805
Simple ANN	**0.813**	**0.833**
Testing on Combined		
RFC	0.804	0.831
SVM	0.776	0.809
Simple ANN	**0.811**	**0.839**

Table 5.16: Results of LOSO evaluations of models trained on the combined dataset. The top part shows the performances of the models on the WESAD dataset, middle part displays the results for ForDigitStress dataset, and the bottom part tabulates the overall performances.

WESAD and ForDigitStress datasets involved social stress (matching type) but had different stress intensities. WESAD induced high-intensity stress, while ForDigitStress participants experienced moderate levels of stress. This finding suggests that stressor type is critical for the generalizability of HRV-based stress detection models.

Although the ForDigitStress dataset was not collected in an industrial environment, it represents a real-world job interview scenario. Moreover, social stress can occur in industrial settings as well. While WESAD doesn't directly reflect a real-world scenario, this study demonstrates that such datasets can still be utilized to train models with broader applicability.

5.6 Reflections and Remarks

This chapter explored the generalizability of stress detection models, focusing on signals that reflect heart activity - ECG and BVP. This exploration was motivated by the scarcity of stress datasets collected in industrial settings. A practical approach in such cases is training stress detection models on publicly available datasets and deploying them in real-world scenarios. However, this approach raises the question: to what extent can these models be applied to the target use case? Identifying the factors impacting the models' applicability enables the development of models compatible with real-world settings.

To address this question, the chapter initially investigated the generalizability of ECG-based deep learning models that typically achieve state-of-the-art performances on the training datasets. For comparison purposes, models based on hand-crafted HRV features were also considered. While the HRV models performed better than the ECG-based models in the generalizability assessment, their generalization capabilities were still limited. Some differentiating factors of the datasets that may have influenced the generalizability of the models were identified.

For further investigations, this chapter presented a social stress dataset. This dataset was leveraged to investigate the generalizability of HRV models trained on datasets with the same type of stressor. Notably, the method of inducing stress and the experienced stress intensity were still different in the datasets. The drastic improvement observed in generalization performances suggests that matching the type of stressor is crucial when the training and deployment scenarios differ.

This chapter limited its exploration to ECG and BVP signals. For a comprehensive assessment of generalizability challenges, these investigations should be extended to other popular modalities used in stress detection, such as EDA.

Part III

Reproducibility and Versatility

Chapter 6

Behaviour Patterns in Industry-like Human-Robot Collaboration

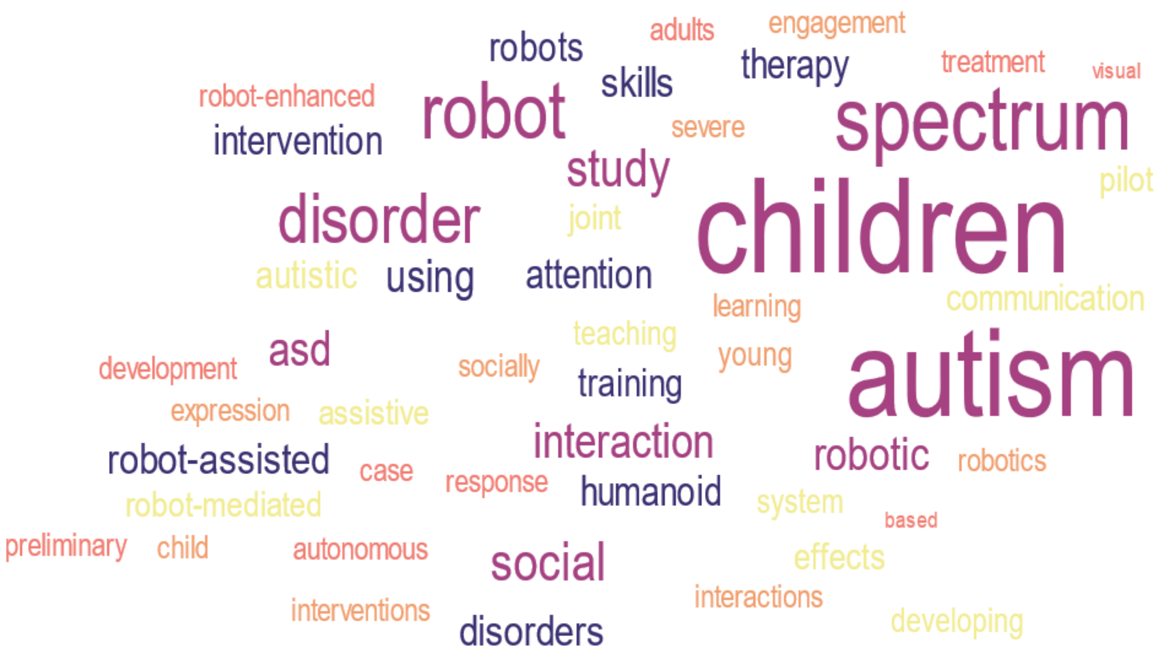


Figure 6.1: A word cloud depicting the prominent topics of Human-Robot Interaction (HRI) research involving Autism Spectrum Disorder (ASD) individuals. The size of a word reflects its frequency in the titles of journal articles retrieved through a literature search. While search terms included "autism" and "robot", the word "children" emerged as a significantly prominent non-query term. This implies that HRI studies predominantly focus on children with ASD.

6.1 Overview

The growing adoption of cobots in various industries has driven research towards ensuring safe and efficient human-robot interaction [Robla-Gómez et al., 2017; Arents et al., 2021; Li et al., 2023]. In small and medium-sized enterprises, this rapid adoption of cobots has led to an emphasis on flexibility and customization [Wadhwa, 2012; Masood and Sonntag, 2020; Kopp et al., 2021]. However, with the advent of Industry 5.0, the emphasize shifted to human-centered cobot adaptations, where cobots are tailored to the specific needs and preferences of operators [Zhang et al., 2023a; Gervasi et al., 2023]. This worker-centric approach is crucial not only for efficiency but also for the mental health and well-being of cobot workers [Xu et al., 2021; Nicora et al., 2021; Lu et al., 2022b].

One particularly important population to consider for cobot integration are the potential workers with ASD. The structured and predictable nature of collaborative cobot tasks aligns well with the strengths of individuals with ASD [Hendricks, 2010; Goris et al., 2020]. Such a workplace can provide supportive environment that fosters their inclusion [Kagermann and Nonaka, 2019; Tomczak, 2021]. However, a brief survey of the literature[1] (topics represented as a word cloud in Figure 6.1) suggests that there is a notable absence of research focusing on industrial Human-Robot Collaboration (HRC) contexts like collaboration, assembly tasks, and handovers.

This chapter presents an industrial HRC scenario involving a collaborative assembly task with the assembly steps divided equally between the human participant and the cobot. The behavioral patterns of the participants were investigated in an exploratory study lasting for a week. Behaviors of eight neurotypical and eight ASD participants were analyzed using quantitative and qualitative tools. The behavioral manifestations of the two groups were also compared to highlight their differences and a potential need for distinct adaptation strategies. The content presented in this chapter was previously published in the following paper and has been expanded here:

* M. Mondellini, P. Prajod, M. L. Nicora, M. Chiappini, E. Micheletti, F. A. Storm, R. Vertechy, E. André, and M. Malosio. Behavioral patterns in robotic collaborative assembly: Comparing neurotypical and autism spectrum disorder participants. *Frontiers in Psychology*, 14, 2023

[*I share first authorship for this paper. I contributed significantly to study design and selection of analysis tools. I performed the video-based analysis for identifying behavioral patterns for both neurotypical and ASD participants. I also contributed significantly to deriving insights based on various outcomes.*]

6.2 Background Literature

6.2.1 Human Factors in Industrial HRC

Goals

One of the primary goals of human-centered Industry 5.0 settings is to improve both the physical and mental well-being of workers [Storm et al., 2022; Xu et al., 2021; Liu and Jebelli, 2022; Verna et al., 2023]. For instance, a cobot moving too fast or too close to the operator without proper adaptations can cause mental fatigue and chronic stress. Works

[1] `https://www.scopus.com/`, Query: TITLE-ABS-KEY ((autism OR asd OR autistic) AND robot* AND (interaction OR collaboration OR coexistence OR colocation OR co*operation))

advocating a shift towards human-centered adaptations in industrial environments highlight additional benefits of such adaptations that lead to improvements in workstation performance and productivity including enhanced efficiency and reduction in errors [Faccio et al., 2023; Simões et al., 2022; Di Pasquale et al., 2023]. For example, a cobot anticipating the operator's intent and actions can respond faster, leading to a more efficient interaction [Huang and Mutlu, 2016; Görür et al., 2018]

Many of the studies in this field focus on improving worker safety and trust/acceptance of the cobot [De Simone et al., 2022; Di Pasquale et al., 2023; Hopko et al., 2022; Baltrusch et al., 2022]. These factors directly influence the willingness of workers to utilize the cobot technology. However, human factors such as engagement, and emotional experience are often overlooked [Storm et al., 2022; Faccio et al., 2023; Toichoa Eyam et al., 2021]. A recent review by Loizaga et al. [2023] explores various human factors critical in Industry 5.0, including aspects like fatigue, attention and stress. Crucially, factors like boredom, stress, fatigue, etc., are major contributors of the variance of human error in manufacturing scenarios [Yeow et al., 2014]. Thus, it is crucial to observe and evaluate which characteristics related to the cobot interaction and which traits of the user may influence these factors.

Operator's Mental Well-being in HRC

Although multiple classification schemes have been proposed for characterizing operator-cobot interaction, a commonly used scheme distinguishes between cooperation and collaboration [Matheson et al., 2019]. Cooperation involves the operator and cobot sharing the same workspace at the same time but working on separate tasks. On the other hand, collaboration involves the operator and cobot working jointly on the same task, requiring real-time coordination. Notably, many works investigating human factors in industrial HRC employ cooperative tasks rather than collaborative tasks [Jahanmahin et al., 2022; Arents et al., 2021]. This distinction is important because the interaction experience and the factors influencing it would differ depending on the level of collaboration.

Several studies have explored how cobot behavior influences mental well-being factors in collaborative tasks. For example, Koppenborg et al. [2017] used virtual reality to investigate the impact of cobot movement speed and trajectory predictability on human operators. Their findings revealed that lower predictability led to decreased performance, while faster cobot movements increased perceived task load and anxiety.

Moving to physical cobot interaction, Gervasi et al. [2022] conducted studies with varying robot configurations, including robot movement speed, operator-robot proximity, and control over task execution time. Analyzing physiological measures and questionnaires, they found that robot movement speed and control over task execution time significantly influenced the interaction quality and stress levels.

Similarly, Su et al. [2024] examined the task load and physiological measures of participants during a collaborative Lego assembly task. The participants interacted with the robot using various methods (button press, hand gestures, and verbal commands). The authors observed that while introducing interaction generally reduced mental stress, the complexity of interactions influenced the extend of stress reduction. Notably, hand gestures elicited higher mental stress than the other two interaction methods.

Toichoa Eyam et al. [2021] proposed a cobot system that adapts its movement speed depending on the operator's mental stress (measured through Electroencephalogram signals). They observed how one operator interacted with this system and demonstrated potential benefits in reducing stress. However, they also observed a slight decrease in engagement levels with repetitions. They interpreted this as the operator becoming used to the task and potentially losing interest over time. This suggests that addressing boredom and monotony alongside stress reduction might be crucial for long-term well-being in collaborative tasks.

The above studies primarily focus on cognitive load and mental stress of the operator and the cobot-related factors that influence them. While these are crucial aspects of industrial HRC, a broader perspective on well-being is necessary. Factors like boredom and distraction can significantly impact task performance and mental well-being in industrial settings. Repetitive tasks, a characteristic of many industrial environments, can lead to monotony and decreased focus, potentially increasing the risk of errors.

Moreover, these studies, and similar research, often utilize short laboratory sessions to study the elicited states [Lu et al., 2022a]. The interactions in these studies lasted for 30 - 90 minutes and were not repeated over days. However, as noted by Toichoa Eyam et al., certain changes in behaviors are observed over time or with repetitions. Furthermore, there is a growing need to replicate industrial conditions in laboratory studies [El Zaatari et al., 2019]. This implies studies should incorporate repetitive tasks and extend over longer periods (e.g., hours of interaction over few days) to mimic real-world industrial conditions.

6.2.2 Involving ASD Participants in Industrial HRC

What is ASD?

ASD is a developmental condition that affects an individual's social interactions, communicative cues, and behavioral patterns [American Psychiatric Association, 2013; Pennazio et al., 2020]. Individuals with ASD experience a wide range of challenges and varying levels of severity. Some commonly observed characteristics include challenges in social skills, repetitive behaviors, and heightened sensory sensitivity [Loucas, 2015]. Manifestations of these characteristics differ from individual to individual. For example, some individuals with ASD have difficulty understanding non-verbal cues during conversations, while some others exhibit repetitive hand movements (e.g., flapping hands).

Research has shown that there is a higher prevalence of ASD in males compared to females [Loomes et al., 2017; Shaw et al., 2020; Adak and Halder, 2017]. While the male-to-female ratio of ASD participants varies depending on the study, a general observation is that ASD is more common in males than females.

ASD Individuals and Industry 5.0

Individuals with ASD are often negatively impacted by traditional hiring practices, which look for qualities like being a team player, communication skills, etc. [Kagermann and Nonaka, 2019]. These skills might not reflect the strengths of individuals with ASD, such as focus, attention to detail, and adherence to routine. While a few works have identified the need for inclusive HRC workplaces for individuals with cognitive disabilities [Kildal et al., 2018; Colombino et al., 2021], there is a lack of research on integrating ASD individuals in such environments.

Many individuals with ASD show significant interest in robots, which can enhance their engagement and motivation in HRI tasks [Scassellati et al., 2012; Raptopoulou et al., 2021]. Studies have shown success in employing social robots to improve therapy outcomes and engagement for children with ASD (e.g., Baraka et al. [2022]; Panceri et al. [2021]). Study by Lytridis et al. [2022] also demonstrates that robot features like LEDs can be effective in engaging ASD children during therapy sessions.

Furthermore, the predictable and routine nature of many industrial HRC tasks aligns well with the preferences of individuals with ASD [Hendricks, 2010; Goris et al., 2020]. Studies like Schadenberg et al. [2021] suggest that predictable robot behavior positively influences attention, which can be beneficial in HRC work environments.

Moreover, the collaborative tasks involved in Industry 5.0 presents an opportunity for inclusion of individuals with ASD. Research like Giraud et al. [2021], which explored interactive systems involving joint activities, demonstrates the potential for technology to foster inclusiveness alongside skill development.

6.2.3 Research Gaps

While research on industrial HRC is constantly evolving, significant gaps remain in understanding the long-term impact on operator well-being and the potential for inclusion of individuals with ASD. These gaps are addressed in the upcoming sections of this chapter.

- Limited understanding of long-term operator well-being: As mentioned in Section 6.2.1, Current research in HRC primarily focuses on mental states like stress and cognitive load, often induced by varying certain parameters like cobot speed or trajectory. These studies are often short-term, lasting for some minutes. While these studies provide valuable insights, they might not capture the cumulative effects of long-term collaboration with robots. Long-duration studies (e.g., sessions lasting for few hours and repeated for some days) are necessary to understand how operator state manifestations and behavioral patterns evolve with multiple repetitions of the task over an extended period.

- Lack of HRC research involving adults with ASD: While the potential of integrating individuals with ASD is discussed in Section 6.2.2, it is important to acknowledge the lack of studies in this direction. As highlighted by Figure 6.1, most existing studies involving robot interaction focus on children with ASD. Moreover, these studies focus on therapeutic applications or skill development rather than investigating how robot behavior can be adapted to facilitate successful collaboration. Hence, there is a need for studies investigating how adults with ASD interact with cobots, especially in collaborative tasks. Understanding these interactions and potential challenges is essential for developing cobot adaptation strategies for worker's with ASD.

6.3 Setup and Data Acquisition

This exploratory study employed a generic collaborative assembly scenario in a laboratory setting to investigate behavioral patterns that emerge during industrial HRC scenarios. To ensure ecological validity, the task design and session duration closely mimicked real-world industrial scenarios. This approach aimed to ensure the collected data reflects natural and representative manifestations of behavioral patterns relevant to actual industrial HRC applications. The study recruited participants who played the role of cobot workers, collaborating with the cobot on the assembly task.

6.3.1 Collaborative Assembly

The task involved assembling a 3D-printed planetary gearbox [Redaelli et al., 2021]. The assembly process was divided equally between the cobot and the human participant.

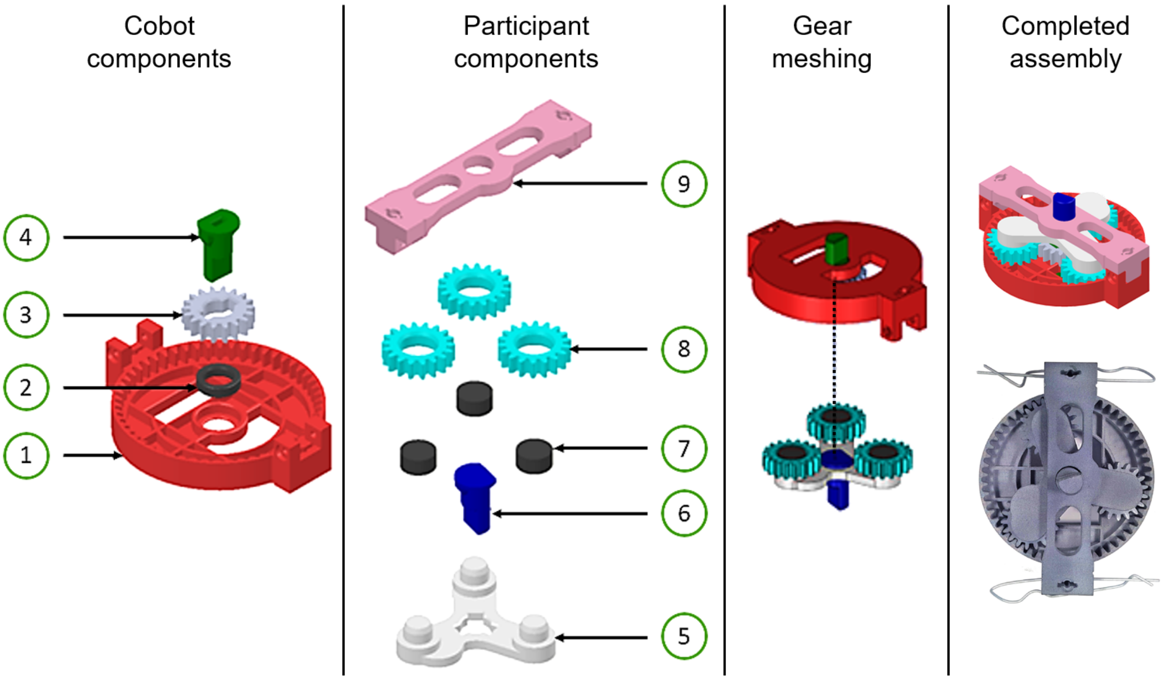


Figure 6.2: An illustration of the individual components used for assembly and the final assembled gearbox resulting from collaborative meshing. The copyright remains with the authors [Mondellini et al., 2023].

Cobot's Tasks

The cobot assembled the first half of the gearbox, i.e., components 1 to 4 in Figure 6.2. The cobot's designated workspace consisted of specific sections on the table in front of it, where the components required for its sub-assembly were placed. To identify and locate these components, the cobot utilized a wrist-mounted Pickit3D camera. Upon successful detection, the cobot picked and placed the components using the Robotiq Hand-e gripper, following a step-by-step process to assemble its sub-assembly (as illustrated in Figure 6.3). If the camera failed to detect a specific component for the next step, the cobot would pause the assembly process. This pause would continue until the missing component became available and an external command was received to resume the assembly process. This error-handling mechanism minimized the production of defective pieces.

Once the cobot completed its sub-assembly, it moved to a designated shared area for joint activity. In this phase, the cobot held its sub-assembly at an angle suitable for meshing the gears with the human participant's sub-assembly, facilitating the collaborative effort. Upon receiving a signal from the participant pressing a foot pedal button, the cobot released its sub-assembly and began a new assembly cycle.

The control architecture integrated ROS [Quigley et al., 2009] and VSM [Gebhard et al., 2012] frameworks. ROS facilitated control of both the detection camera and the cobot, while VSM enabled the execution of the programmed assembly steps by the cobot.

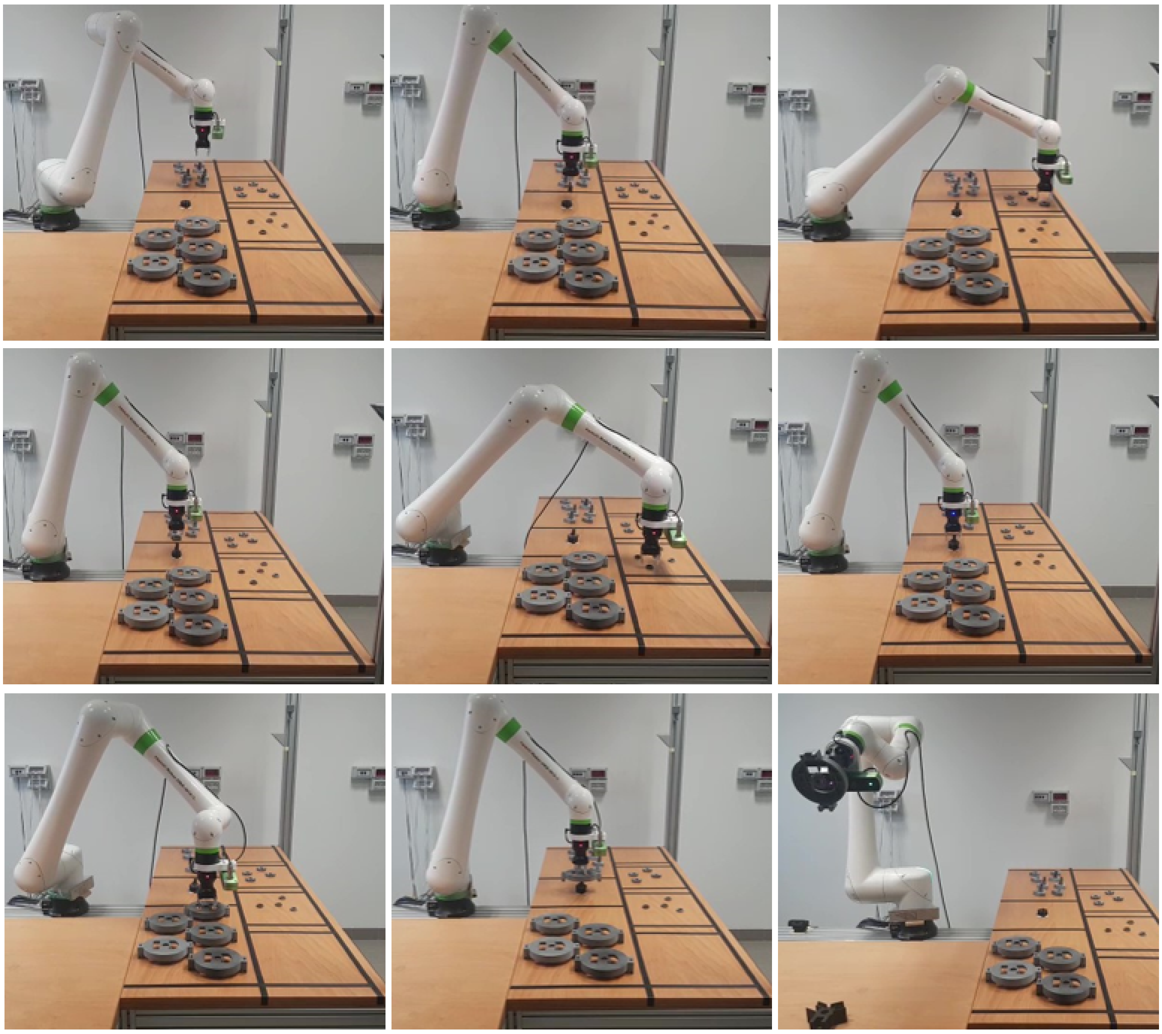


Figure 6.3: A sequence of snaps from a video showing the various steps of the cobot's assembly process.

Participant's Tasks

The human participant assembled the remaining components (5-8 in Figure 6.2). If needed, they utilized a support structure specially designed for this task to aid them in their assembly steps. Once both sub-assemblies were complete, they collaborated with the cobot to join the two sub-assemblies by meshing the gears. After successful meshing, the participant triggered the cobot to release the meshed sub-assemblies. They then completed the assembly by covering the product using component 9, resulting in the final product (see Figure 6.2).

In addition to the assembly, the participants were responsible for managing the components involved in the entire assembly process. They were provided with a box containing all the components needed for both themselves and the cobot. Importantly, they ensured the cobot's table was sufficiently stocked with spare parts by replenishing any low-running components from a nearby box.

The task design provided participants with a high level of autonomy through an unconstrained assembly process. This meant they could assemble as many sub-assemblies as possible, as long as they had the necessary components. They were also not restricted in the timing of their component replenishment or assembly steps, allowing them to work at their own pace. Furthermore, the experiment did not impose any production targets.

6.3.2 Layout

Figure 6.4 illustrates the layout of the experimental setup. The L-shaped table served as the primary workspace for the assembly activities. The participant utilized the participant's side of the table for assembling their sub-assembly. This side was equipped with a fixed, support structure designed to assist participants in their sub-assembly steps. A camera was positioned approximately 1.5 meters from the participant on a separate support structure to capture their upper body during the assembly process.

The other side of the table served as the cobot's workspace, designated for its sub-assembly activities. Components for the cobot's sub-assembly were placed on this side. The cobot performed its assembly at a fixed location on this side of the table. A designated shared space at the junction of the tables facilitated the collaborative joining activity between the participant and the cobot. A Fanuc CRX10iA/L collaborative robot was positioned at the corner of the L-shaped table configuration, ensuring its reach to both the cobot's workspace and the collaboration space.

As mentioned before, the participants were provided with all the components in a box. Additionally, an empty box was provided to store the completed assemblies. They were free to choose the placement of these boxes. Some of the participants utilized a part of their table to keep the boxes, whereas the others placed them on the ground.

6.3.3 Participants

This study involved 16 participants, belonging to two groups:

- **Neurotypical (NT) group**: 8 participants (5 females, 3 males) belonging to the age range of 18-30 years.
- **Autism Spectrum Disorder (ASD) group**: 8 high-functioning individuals (1 female, 7 males), aged between 21 and 50 years old. Notably, all participants in this group had an IQ exceeding 70, confirming the absence of intellectual disability.

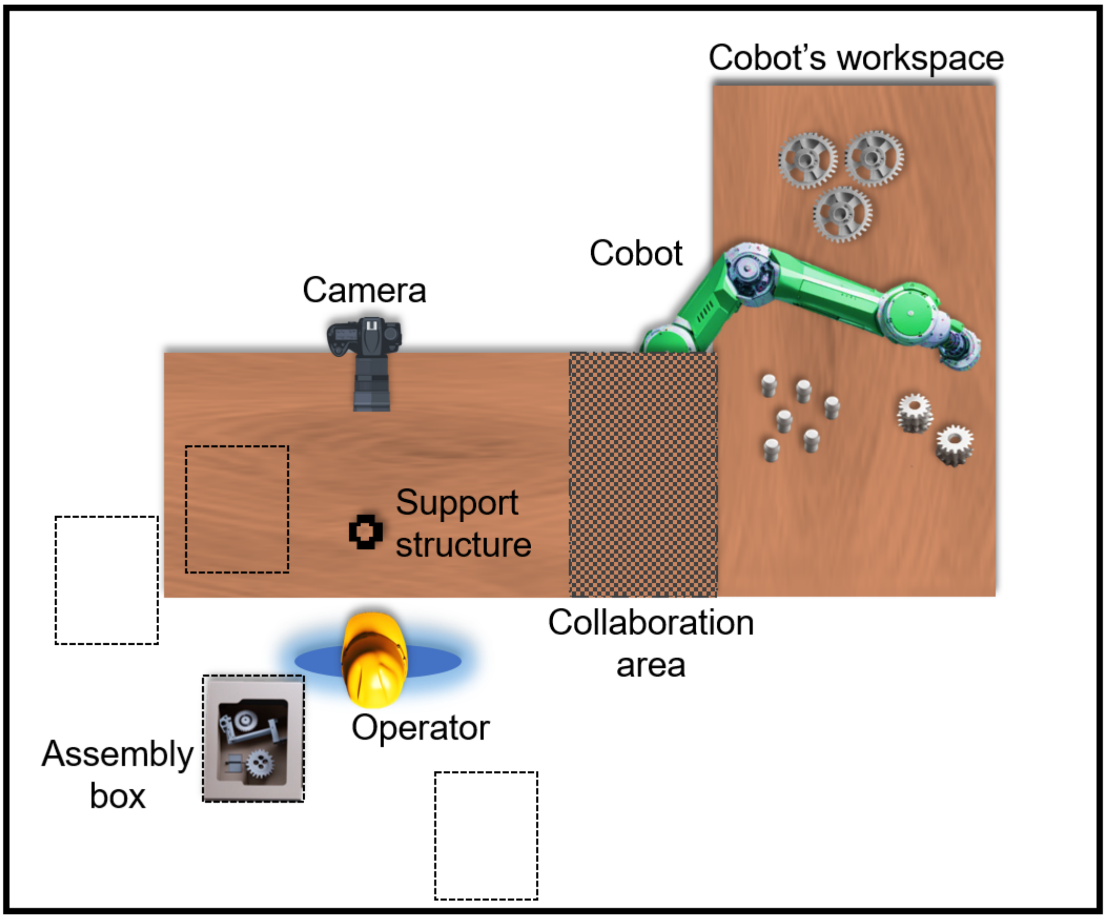


Figure 6.4: An overview of the layout of the experimental setup showing the participant's table, cobot's table, and collaboration area. The tables were set in an L-shaped configuration, and the cobot was positioned at the junction. A camera facing the participant captured video snippets of the collaborative assembly. Participants kept the assembly box (containing components) within reach on their table. The dotted lines illustrate some of the observed box placement variations.

It is important to acknowledge the gender imbalance towards males in the ASD group. However, based on previous studies, gender imbalances are expected in ASD groups [Loomes et al., 2017; Shaw et al., 2020; Adak and Halder, 2017]. Furthermore, none of the participants had prior experience working in an industrial setting with a cobot.

Participants were engaged in the study task for 3.5 hours daily for five consecutive days, spanning Monday to Friday. This extended duration aimed to capture and analyze potential modifications in performance and behavior over time as they familiarized themselves with the task and the cobot. Considering the length of the sessions and duration of the study, participants were recruited based on their availability to travel independently to the study location (by car or train) or to stay in a nearby facility for the entire week. Additionally, to facilitate participation and minimize potential discomfort for individuals with ASD, they received comprehensive briefings before commencing the experiment week. These briefings covered the people they might encounter, the specific tasks they would perform, and the daily routines of the lab (e.g., security protocols, break times, etc.).

The study took place at National Research Council of Italy - Lecco campus. The neurotypical participants were recruited from a nearby institute and the ASD participants were contacted through Auticon. This study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the Ethics Committee of I.R.C.C.S. Eugenio Medea (protocol code N. 19/20—CE of 20 April 2020).

6.3.4 Video Recordings

A Logitech C920 Pro HD webcam was positioned in front of the participants (see Figure 6.4) to capture their behavior during the experiment. Videos were recorded in HD format (1280x720) at a frame rate of 25 fps. The upper half of the participant's body was visible in the recordings. Considering the participant's need to move around the workspace during the assembly process, the camera was strategically positioned to keep the participant in the camera's field of view for the maximum duration.

To capture potential changes in behavior over time, three 10-minute video recordings were taken for each participant on the first and last days of the experiment. These recordings were captured at the beginning, middle, and end of each workday, resulting in a total of one hour of video data per participant. This yielded a total of 16 hours of video recordings for analysis. These video recordings serve as the primary data source for both the qualitative and quantitative analyses described in the subsequent sections.

6.4 Analysis Methods

Given the limited understanding of behavioral patterns in industrial HRC settings, especially for individuals with ASD, this study adopted a multi-tool analysis approach. This approach utilizes four distinct tools to capture behaviors during both predictable and unpredictable instances. Two of the chosen tools, like video annotations, are designed for the precise observation of predefined aspects of collaboration, including elements such as gaze patterns and hand movements. Conversely, the other two tools, like live note-taking, facilitate capturing responses to unpredictable scenarios in a long session. In addition to enabling a qualitative exploration of observed behaviors within each group, the chosen tools also allow for a quantitative comparison between the neurotypical and ASD participants. It is important to acknowledge that some of the quantitative measures were tailored to the study's assembly task, and computed for comparative analysis within the study context. These measures should not be interpreted as broader generalizations about the participants' overall efficiency or effectiveness.

6.4.1 Observational Grid

To capture specific predefined aspects of collaboration, an observational grid was employed. This grid, inspired by Roller and Lavrakas [2015], enabled systematic recording and categorization of specific observable behaviors relevant to the participants' collaboration experience. As this method is particularly effective for shorter experimental sessions, it was well-suited for analyzing the collected video data.

The observational grid was initially designed with four categories to capture specific aspects of the participants' behavior:

1. **Manifestations of tiredness**: This category aimed to identify body movements or facial expressions indicating fatigue or tiredness, and potential differences in how participants from both groups expressed it during the task. This category was chosen specifically in the light of growing literature highlighting the importance of detecting tiredness at the workplace [Åkerstedt et al., 2004; Sadeghniiat-Haghighi and Yazdi, 2015; Gabriel et al., 2018]. Since boredom due to the repetitive nature of the task can also lead to fatigue [Caldwell et al., 2019], manifestations of boredom were also included in this category.

2. **Hand gestures**: This category recorded hand movements that were not related to the task such as touching the nose, flapping hands, etc. This information was used to investigate whether participants with ASD exhibited distinct patterns of hand gestures compared to the neurotypical group, similar to observations in other contexts [Goldman et al., 2009; Grossi et al., 2021].

3. **Assembly methods**: This category focused on how participants assembled the gearbox, including aspects like one-handed vs. two-handed assembly or building multiple pieces simultaneously. This data facilitated the exploration of potential differences in adherence to routine and behavioral rigidity between the groups, as individuals with ASD are often known to be inflexible during repetitive activities [D'Cruz et al., 2013; Poljac et al., 2017; Petrolini et al., 2023].

4. **Loading cobot table**: This category recorded the participant's timing of adding new components to the cobot's table (e.g., when the cobot stops, after finishing a gearbox, at any time). This category also aimed to investigate potential differences in behavioral rigidity in the two groups.

After an initial viewing of video recordings, additional categories were deemed essential for understanding participants' behavior. Hence, the grid was expanded to include:

5. **Other manifestations**: This category captured behaviors unrelated to fatigue but contributing to understanding the participants' state, such as fanning oneself due to heat.

6. **Regard for cobot**: This category recorded participant reactions towards the cobot's actions, including staring, talking, or even ignoring the cobot's cues for joint activity. This category was added because existing literature suggests social robots are especially engaging for individuals with ASD Pennisi et al. [2016]; Kumazaki et al. [2020].

7. **Talk to someone**: This category noted instances where participants interacted verbally with others in the room (e.g. experimenter).

Finally, a dedicated space for “Notes" was included to record any additional observations during video analysis. While not aiming to comprehensively categorize participants' behaviors, this grid effectively captured recurring patterns relevant to the industrial HRC experience. One of the researchers, who has a psychology background, completed the observational grid based on the videos. An example of the final grid with data from one participant is presented in Table 6.1. The data from the observational grid was utilized for qualitative analysis of the observed behavioral patterns.

Category	Entry
ID	4014006
DAY	Day 1 Video 2
MANIFESTATION OF TIREDNESS	Participant looks at the clock (2 times)
HAND GESTURES	Scratch the nose (1 time), Scrub hands (1 time)
ASSEMBLY METHODS	–
LOADING COBOT TABLE	–
OTHER MANIFESTATIONS	Tight lips (1 time), Wet mouth with tongue (5 times)
REGARD FOR COBOT	Cobot arrives, participant prefers to finish assembling all their sub-assemblies
TALK TO SOMEONE	yes
NOTES	Rubs hands after completing action, plausibly showing satisfaction

Table 6.1: An example of the completed observational grid, illustrating recorded observations from a single video recording of a participant.

6.4.2 Live Note-taking

Building on the understanding that ASD diagnosis relies on behavioral markers [American Psychiatric Association, 2013] and that individuals with ASD exhibit unique and potentially diverse behavioral patterns, unstructured live note-taking was employed alongside the observational grid as a qualitative measure. This additional data collection method aimed to minimize the potential loss of relevant behavioral information not captured by the predefined categories of the grid.

Two researchers, separate from those who utilized the grid, observed participants non-intrusively for 3.5-hour sessions on 3 days spread across the week (Monday, Wednesday, and Friday). One note-taker has psychology background and the other is themselves diagnosed with ASD. Both of them had previously worked with ASD participants. Unstructured notes were taken without a predetermined framework, enabling the recording of unforeseen behaviors that might occur during the interaction between the ASD participants and the cobot.

These notes were then analyzed qualitatively using an “empathy map" approach [Nielsen Norman Group, 2018], which visualizes an individuals profile in terms of “Says", “Thinks", “Does", and “Feels". An adapted version of empathy maps were utilized to form informative cards named “persona cards" (see Table 6.2 for an example). These personas summarize each ASD participant’s profile across five key categories:

1. **Task**: This category highlights the participant’s challenges and strengths in interacting with the cobot during different phases of the assembly task. Examples include managing cobot stops, using the pedals effectively, and maintaining focus on the task.

2. **Work Organization**: This category describes the participant’s strategies for organizing their work, such as managing the supply of components to the cobot or performing multiple tasks simultaneously.

3. **Say - Quotes**: This category captures verbalizations uttered by the participant during the assembly task.

4. **Act - Recurrent Behaviors**: This category describes repetitive behaviors not directly related to the task, such as checking a phone, crossing arms, or snapping fingers.

5. **Feel - Emotional Expressions**: This category captures any observed emotional expressions during the experiment, such as smiling or singing.

The Say - Quotes reflects the Says category in empathy map, Feel - Emotional Expressions reflects Feels category, and the other three reflect the nuances of Does category.

It is important to note that due to the unstructured nature of the data, only qualitative descriptions of behaviors were possible, not frequency quantification. Additionally, the primary objective of live note-taking was to capture novel aspects of the long-term interaction between cobots and individuals with ASD. So, this measure was limited to the ASD group and therefore a comparison between groups was not possible.

6.4.3 NOVA Annotations

This study employed NOVA [Baur et al., 2013], a tool for annotating and analyzing behaviors in social interactions. NOVA’s graphical interface facilitates the annotation of multimodal data from various sources like video, audio, and bio-signals. Similar to the observational grid, NOVA is well-suited for analyzing short experimental sessions.

In this study, NOVA enabled the quantitative analysis of videos recorded by the frontal camera using frame-wise labeling. This allows researchers to mark specific moments and

Person ID	
Task Challenges	
Needs help to stop the cobot	Moves components around too much
Task Strengths	
Works close to the cobot	Can operate system correctly
Quick movements	
Work Organization	
Fills buffer while cobot moves	Empties box and organizes parts
Works on two sub-assemblies in parallel	Adds parts if detection fails
Takes break autonomously	
Say - Quotes	
“Oh no!" (no components for cobot)	“Where does this noise come from?"
“What happens with these pieces?" (after the shift)	“The gripper is behaving in a weird way"
Act - Recurrent Behaviors	
Some parts fall due to quick movements	Stretch their back
Feel - Emotional Expressions	
No visible fear of cobot	

Table 6.2: An illustrative example of a participant profile or persona card summary based on live note-taking.

label different participant behaviors within the video frames. Additionally, NOVA’s interface can handle data from multiple individuals, facilitating the analysis of interactions between different entities like the participants and the cobot.

Beyond annotation and visualization capabilities, NOVA offers exporting annotations in popular formats like Excel. The exported annotations include start time, end time, and label for each identified behavior. This allows further analysis of the annotated data.

In this study, the NOVA tool was used to quantitatively analyze specific actions performed by participants and compare the neurotypical and ASD groups. The process involved creating two separate annotation tracks within NOVA: one for the participant and one for the cobot (as depicted in Figure 6.5). Based on initial viewing of the videos, the following labels were identified by the researchers along with what actions correspond to each label. One of the researchers, who is familiar with affective signals and NOVA tool, annotated the videos. A default label called “Other" was utilized to denote parts of the video that were difficult to classify (e.g., blurry, transitions between phases of assembly, etc.). This annotation enabled the measurement of the duration of specific actions for each entity within the video frames.

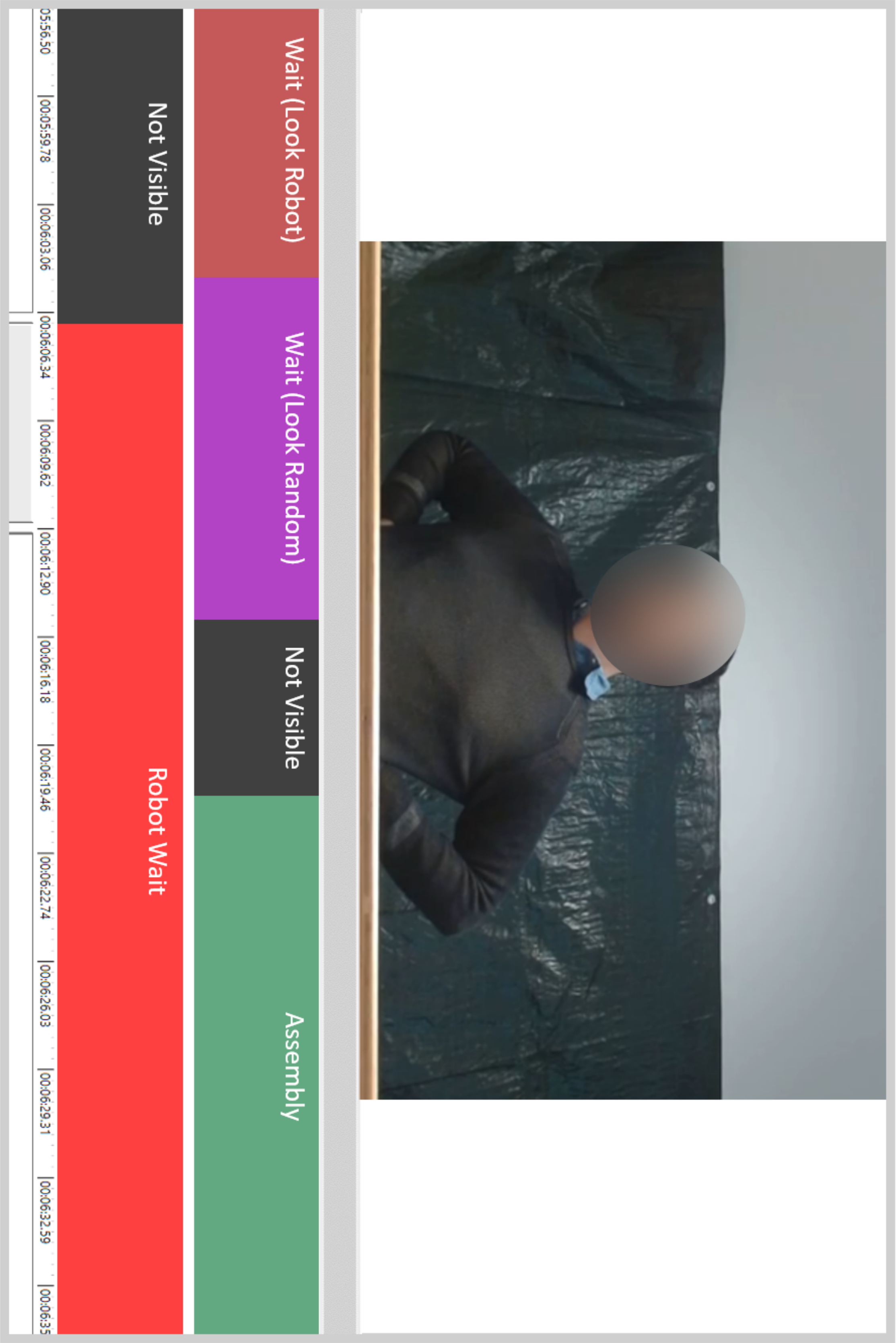


Figure 6.5: An example visualization of NOVA video annotation for a participant. The image shows a video frame of the participant looking at the cobot. The top track has the labels for the participant's actions and the one below has labels for the robot's action. The copyright remains with the authors [Mondellini et al., 2023].

Participant's Labels

The annotation scheme focused on three primary participant actions:

1. **Gathering**: This label marked instances where the participant collected the necessary components for the assembly.

2. **Assembling**: This label denoted the time participants spent building a sub-assembly from the gathered components.

3. **Final Joining**: This label identified periods where the participant meshed the gears of their sub-assembly with the cobot's sub-assembly to form the final product. This is also called the joint activity phase of the production cycle.

Additionally, the annotation scheme incorporated labels to differentiate between two types of waiting behaviors exhibited by participants:

4. **Wait (Look Robot)**: This label represented instances where the participant maintained their gaze on the cobot while they were not actively engaged in any of the assembly steps. This behavior indicates that they were potentially waiting for the cobot to complete its actions.

5. **Wait (Look Random)**: This label signified periods when the participant waited but engaged in other activities, such as looking around the environment, talking to someone else, or exhibiting other forms of distraction.

Furthermore, the label "**Not Visible**" was utilized to account for the missing data during the occasional moments when the participants moved outside the camera's field of view.

Cobot's Labels

The tip of the cobot's wrist was occasionally visible in the video, especially when it was at the collaboration area for the joint activity. So, three labels were included in the annotation scheme describing the cobot's actions:

1. **Robot Wait**: This label captured the instances where the cobot arrived at the collaboration area and waited for the participant to perform the joint activity. The duration of this action was directly linked to the participant's pace and their decision-making regarding the timing of the joint activity.

2. **Final Joining**: This label signifies the cobot's participation in the joint activity, aligning with the participant's "Final Joining" label.

3. **Not Visible**: This is the default label, denoting that the data regarding the cobot's actions were not available.

A key difference between observational grid and NOVA annotation is the origin of the coding scheme. The observational grid was designed by taking into account known behavioral patterns in ASD individuals, whereas NOVA labels were derived from the various phases of an assembly cycle that are not specific to either of the groups. Moreover, the observational grid identifies the occurences of behaviors and how they change over the week, while NOVA measures the duration of specific actions in the assembly cycle.

6.4.4 Week-long Performance

While the previously mentioned tools focused on the participant's behavioral aspects during the collaborative assembly task, additional data was collected to quantify their overall performance throughout the week.

An Excel spreadsheet was used to record the following information for each participant every day of the experimental week:

1. **Start and end time of the session**: This enabled the computation of the total duration of the daily experimental session.

2. **Activity Stops**: Any instances where the assembly activity was paused or interrupted were documented.

3. **Number of assembled gearboxes**: The total number of gearboxes assembled each day was recorded.

This data was used to compute the daily "**Up-time**", which refers to the total active working time of the participant. This duration excluded breaks requested by the participants and unexpected interruptions due to factors like cobot malfunctions. A performance index was then computed for each participant across the entire experiment week. This index represented the ratio between the daily number of completed assemblies and the corresponding daily up-time. Additionally, the influence of downtime on the performance was also analyzed.

6.5 Qualitative Analysis

The qualitative analysis presented in this section delves into the behavioral patterns observed during the study. As previously mentioned, the observational grid served as a valuable tool for tracking the predefined behavioral constructs within each group. This information, presented as the number of participants (N) exhibiting each behavior out of the total group size (eight individuals in each group), forms the foundation for a qualitative comparison between the neurotypical and ASD groups. Additionally, insights from persona cards summarizing the behavioral profiles of the ASD participants are presented.

6.5.1 Grid Patterns in Neurotypical Participants

This section presents the behavioral patterns of neurotypical participants, observed using the observational grid. The analysis explores how these behaviors manifested and changed throughout the week, comparing the first and last days of the experiment. The prominent patterns are summarized in Tables 6.3, 6.4, and 6.5.

Manifestations of Tiredness

On the first day, several participants exhibited behaviors suggestive of tiredness, including leaning on the table while waiting, and stretching. Manifestations of boredom such as hand activities potentially used to occupy themselves (e.g. fidgeting, tapping fingers on the table) were observed. Additionally, time monitoring behaviors (e.g., looking at the clock, checking phones) were also observed. The number of participants exhibiting these behaviors increased on the last day. Moreover, the frequency of these behaviors in the same individual increased on the last day, potentially due to longer waiting periods. Notably, only one person sat down on both days, albeit with an increased frequency on the last day.

Hand Gestures

Hand gestures, such as touching hair, face, and glasses, were observed throughout the week. There was no significant change in the variety of these behaviors, but an increase in frequency on the last day was observed. This suggests these gestures might be habitual and independent of the task itself.

Assembly Methods

Initially, diverse assembly approaches were observed. Three participants assembled as they retrieved the components, while one participant emptied the entire box first. Three participants followed a sequential assembly strategy, but one participant switched in between to a parallel assembly strategy. However, by the last day, most participants adopted a parallel assembly strategy, suggesting adaptation and potentially improved efficiency. Notably, some participants demonstrated advanced skills, such as multitasking by assembling a new gearbox while waiting with another sub-assembly in hand.

Loading Cobot Table

The first-day observations revealed that three participants preferred to fill the cobot table only when it had the components for one sub-assembly. So, even small delays in restocking led to pauses in the assembly activity. One participant shifted the components on the cobot table causing the cobot to stop due to misplaced components. By the end of the week, the number of such errors decreased, indicating improved awareness and performance.

Other Manifestations

On the first day, only two participants displayed heat-related behaviors (waving hands/shirt). On the last day, one participant was seen humming. Since the study took place in a different week for each participant, it is difficult to draw inferences from these observations.

Regard for Cobot

While most participants focused on the cobot, a few exceptions were observed. On both the first and last days, one participant did not look at the cobot during assembly and occasionally seemed unaware of the cobot's waiting state. In one instance, a participant did not have the required sub-assembly prepared when the cobot arrived for the joint activity. In another instance, a participant prioritized emptying the box over the joint activity. However, these were isolated instances and not frequent behaviors.

It is important to note that on both days, all participants generally exhibited gaze behavior directed at the cobot while waiting.

Talk to Someone

The number of participants who talked to others during the experiment increased from two on the first day to four on the last day.

Inferences

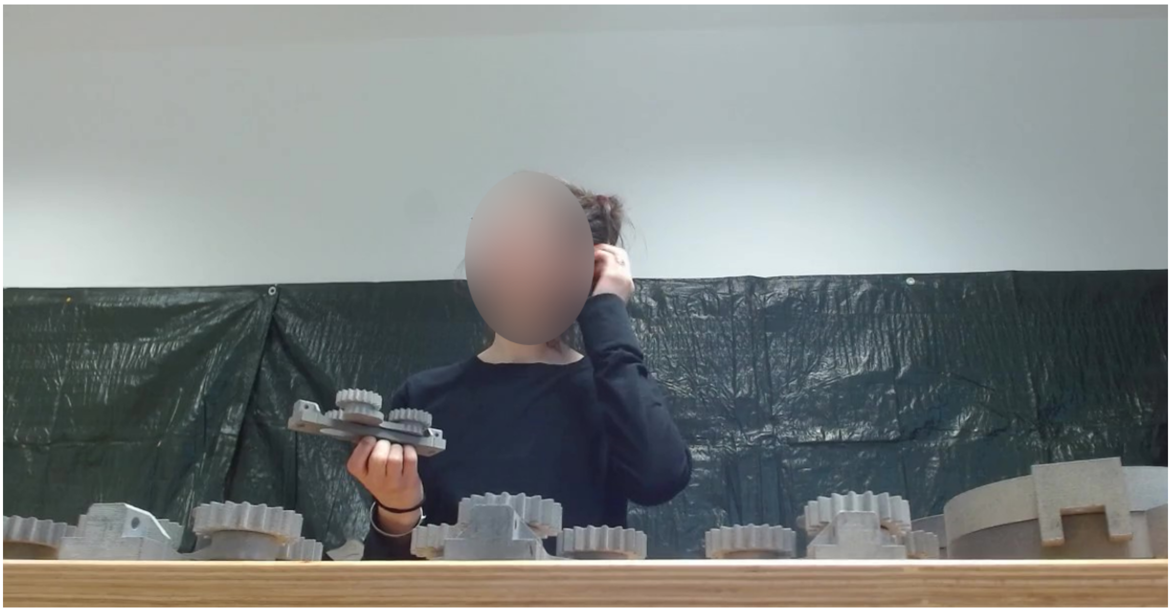


Figure 6.6: A screenshot of a neurotypical participant's workspace during the experiment. Several sub-assemblies are already completed and visible on the table. The participant is also looking at the cobot while waiting, a frequent behavior observed in the neurotypical group. The copyright remains with the authors [Mondellini et al., 2023].

These observations suggest that participants adapted their behavior over time, potentially due to increased familiarity with the task and the cobot. Notably, a shift towards more efficient assembly methods and reduced errors were observed. Once the participants became accustomed to the task, they began assembling multiple sub-assemblies for future use on the table, as illustrated in Figure 6.6.

Moreover, all the participants completed their assembly tasks faster than the cobot, resulting in a considerable amount of waiting time. Observations suggest that the waiting periods resulted in a possible increase in fatigue and boredom over time, leading to the emergence of individual coping mechanisms.

First Day	Last Day
Manifestation of Tiredness	
Hands/arms on table while waiting for cobot (N=5)	Hands/arms on table while waiting for cobot (N=8)
Movements of hands (N=3)	Yawn (N=2)
Hands on hips (N=2)	Snort (N=1)
Sit (N=1)	Sit (N=1)
Time monitoring (N=3)	Time monitoring (N=4)
Stretch (N=2)	Stretch (N=1)
	Fidgeting (N=3)
	Playing with clips (N=1)
Hand Gestures	
Rub hands/fingertips (N=3)	
Rub face (N=4)	Rub face (N=6)
Touch hair (N=3)	Touch hair (N=3)
Adjust clothes (N=1)	
Touch glasses/watch (N=1)	Touch glasses/watch (N=4)

Table 6.3: The observed behaviors in the neurotypical group, summarizing the "Manifestation of Tiredness" and "Hand Gestures" categories of the observational grid

6.5.2 Grid Patterns in ASD Participants

The behavioral patterns from the observational grid for the ASD group are presented in Tables 6.6, 6.7, and 6.8. Below is a brief discussion of the observed behaviors.

Manifestations of Tiredness

On the first day, participants exhibited behaviors indicative of tiredness, such as placing hands/arms on the table while waiting, crossing arms, and hands resting on hips. Other behaviors such as stretching, sighing, and yawning were also observed, but less frequently. Time monitoring behaviors were also observed, with three participants checking the clock and one using their phone to look at the time (8 times in 3 videos). For the same participant, the frequency of these manifestations increased over the day, suggesting an accumulation of tiredness and boredom.

Similar behaviors like resting hands/arms on the table, monitoring time, yawning, stretching, etc., were observed on the last day. Notably, unlike the first day, the frequency of these manifestations remained similar throughout the last day.

First Day	Last Day
Assembly Method	
Assemble as components are taken out (N=3)	
Empty box before the assembly (N=1) - strategy changed	
Sequential assembly (N=3), (N=1) changed strategy	
Parallel assembly (N=6)	Parallel assembly (N=7)
Use of support structure (N=3)	Use of support structure (N=3)
Loading Cobot Table	
Last set on cobot's table (N=3)	
Add components as soon as consumed (N=1)	Move component after placing on table - causes error (N=1)

Table 6.4: The observed behaviors in the neurotypical group, summarizing the "Assembly Method" and "Loading Cobot Table" categories of the observational grid

First Day	Last Day
Other Manifestations	
Manifestation of heat (N=2)	Hum (N=1)
Regard for Cobot	
React in advance (N=1)	
Low awareness of cobot wait (N=1)	Low awareness of cobot wait (N=1)
Look at cobot while waiting (N=8)	Look to cobot while waiting (N=8)
Talk to Someone	
N=2	N=4

Table 6.5: The observed behaviors in the neurotypical group, summarizing the "Other Manifestations", "Regard for Cobot", and "Talk to Someone" categories of the observational grid

Hand Gestures

Three participants frequently rubbed their hands, with one participant focusing on fingertips and another celebrating completion by clapping their hands after rubbing. Four participants exhibited frequent face rubbing throughout the recordings. Some individual behaviors were also observed such as touching glasses, repetitive stereotypical movements, and shaking the wrist with the watch. One participant moved the box around multiple times before settling on a position.

Participants generally displayed consistent and preferred hand gestures on both days. Face touching emerged as the most common hand gesture observed in the ASD group.

Assembly Methods

While most participants emptied the entire box of components before starting assembly, only two participants began assembling as they took the components out. Several participants worked on one sub-assembly at a time, whereas only three participants adopted a parallel assembly strategy. One participant achieved faster assembly by skipping the dedicated assembly support structure, while another arranged all components close together.

On the last day, most participants maintained their initial assembly methods. Only one participant shifted from a sequential to a more efficient parallel approach. While participants showed adaptation in terms of speed, they generally struggled to transition to more efficient assembly methods.

Loading Cobot Table

Participants did not exhibit any specific strategies for loading pieces onto the cobot table. An exception was observed with one participant who adopted a proactive approach of adding new components immediately after the cobot finished a sub-assembly. This strategy ensured that components required for at least two sub-assemblies were always available on the cobot table, minimizing unexpected cobot stops. The loading strategies remained consistent throughout the experiment for all participants.

Other Manifestations

The ASD participants showed expressions of effort (frowning, lip pursing, etc.) when they encountered difficulties in assembly. One participant waved at the camera on the first day. In the videos from the first day, many participants also showed manifestations of heat and frequent wetting of lips. On the last day, some non-task-related movements like body swaying, jumping in place, etc., were observed.

Regard for Cobot

A common behavior was prioritizing completing sub-tasks like emptying the box or preparing new assemblies, even when sub-assemblies were ready for the joint activity, causing the cobot to wait. Two participants frequently looked at the cobot during assembly and waiting periods, possibly to monitor its arrival for collaboration. Two participants did not adjust their assembly pace to the cobot's timing. They would watch the cobot assemble its half, but not have their sub-assemblies prepared for the joint activity.

Three participants displayed facial expressions in response to the cobot's actions, including astonishment and disappointment towards the cobot's speed or errors.

On the last day, one participant showed significant improvement by learning to anticipate the cobot. This led to the participant picking up their sub-assembly a few seconds before the cobot arrived for collaboration. Furthermore, a tendency to reduce the amount of time the cobot waited was observed in a few participants.

Talk to Someone

On the first day, one participant engaged in conversation with another person in the room. This behavior was observed again on the last day, but with a different participant. Overall, the observations related to participants talking to others were minimal.

Inferences

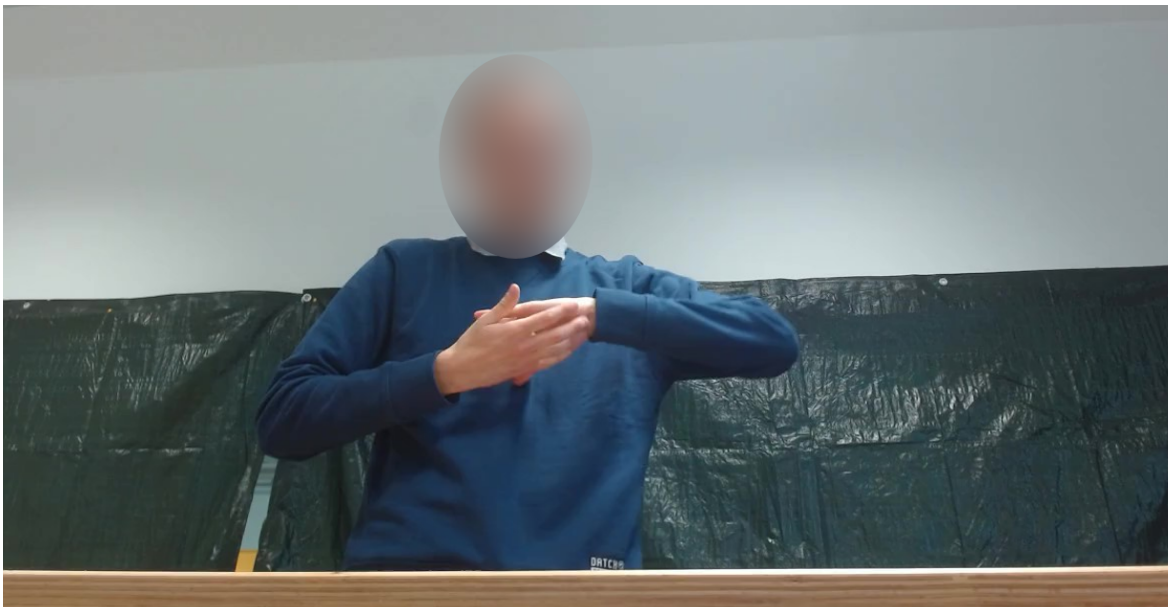


Figure 6.7: A screenshot of an ASD participant's workspace during the experiment. The table is empty, with no sub-assemblies prepared in advance. The participant is also exhibiting a hand gesture (rubbing hands). The copyright remains with the authors [Mondellini et al., 2023].

The ASD participants exhibited a range of behaviors while working with the cobot. There were many manifestations of tiredness and boredom. This suggests potential areas for improvement in task design for better engagement. They also showed some repetitive hand gestures and other body movements typically associated with ASD individuals.

Notably, they tend to maintain their assembly routines. This implied that some participants who started with a “one assembly at a time" strategy (see Figure 6.7), followed the same strategy throughout the week.

Additionally, participants exhibited varying levels of regard for the cobot. Some failed to synchronize their actions with the cobot's joint activity timing.

6.5.3 Grid Comparison Insights

The following key behavioral differences were observed between participants with ASD and neurotypical participants while comparing the summarized observational grids:

- **Earlier signs of tiredness and boredom**: Compared to the neurotypical group, participants with ASD showed signs of tiredness and boredom earlier and more frequently. This might be evidenced by behaviors like looking at their watches while the cobot was working.

First Day	Last Day
Manifestation of Tiredness	
Hands/arms on table while waiting for cobot (N=4)	Hands/arms on table while waiting for cobot (N=3)
Arms crossed (N=2)	Arms crossed (N=2)
Hands on hips (N=1)	Hands on hips (N=3)
Sit (N=1)	Sit (N=1)
Time monitoring (N=4)	Time monitoring (N=5)
Stretch (N=1)	Stretch (N=3)
Yawn (N=1)	Yawn (N=1)
Sigh (N=1)	Close eyes (N=1)
Hand Gestures	
Rub hands/fingertips (N=4)	Rub hands/fingertips (N=4)
Rub face (N=4)	Rub face (N=4)
Clap hands (N=1)	Push components (N=1)
Shake wrist (N=1)	
Stereotypical movements (N=1)	
Move box around before fixing location (N=1)	
Touch glasses (N=1)	Touch glasses (N=2)

Table 6.6: The observed behaviors in the ASD group, summarizing the "Manifestation of Tiredness" and "Hand Gestures" categories of the observational grid

- **Stereotyped hand movements**: The ASD group displayed more repetitive hand movements, such as rubbing their fingertips or hands together. This contrasts with the neurotypical participants, who used a wider variety of hand gestures.

- **Different adaptation rates**: While both groups used similar methods to assemble the parts, the neurotypical participants adapted their strategies much faster, particularly regarding the sequence, timing, and positioning of actions. The ASD group, on the other hand, tended to maintain their initial strategies.

- **Difficulty with multitasking**: The ASD group demonstrated a higher frequency of adopting a sequential assembly approach compared to the neurotypical group, which tended to favor parallel assembly methods. This could be attributed to known challenges with multitasking in ASD [Yang et al., 2017; Mackinlay et al., 2006].

- **Prioritizing own tasks**: ASD participants often continued with their ongoing tasks even when the cobot was ready for the joint activity. In contrast, neurotypical participants prioritized the cobot, leading to less waiting time for the cobot.

First Day	Last Day
Assembly Method	
Assemble as components are taken out (N=2)	
Empty box before the assembly (N=6)	
Sequential assembly (N=4)	Sequential assembly (N=4), later one changes strategy
Parallel assembly (N=3)	Parallel assembly (N=3/4)
No support structure (N=1)	No support structure (N=1)
Components placed close (N=1)	
Loading Cobot Table	
Add components as soon as consumed (N=1)	Add components as soon as consumed (N=1)

Table 6.7: The observed behaviors in the ASD group, summarizing the "Assembly Method" and "Loading Cobot Table" categories of the observational grid

First Day	Last Day
Other Manifestations	
Manifestation of effort (N=2)	Manifestation of effort (N=1)
Greet camera (N=1)	Jump in place (N=1)
Manifestation of heat (N=2)	Sway body (N=1)
Wetting lips (N=3)	
Regard for Cobot	
Make cobot wait (N=5)	Make cobot wait (N=3)
Look at cobot (N=2)	
Facial expressions responses (N=3)	
Watch cobot without doing own assembly (N=2)	
React in advance (N=1)	
Talk to Someone	
N=1	N=1

Table 6.8: The observed behaviors in the ASD group, summarizing the "Other Manifestations", "Regard for Cobot", and "Talk to Someone" categories of the observational grid

- **Lower utilization of monitoring information**: While ASD participants sometimes looked at the cobot, this gaze did not necessarily translate into adapting their actions to collaborate effectively. In contrast, the neurotypical group looked at the cobot more strategically. This could be because they were waiting for the cobot or trying to better time their own assembly steps.

- **Preference for personal space**: Some participants with ASD preferred to maintain a greater distance from the cobot throughout the sessions. This preference was particularly evident during the timing of loading components onto the cobot's table. Unlike the neurotypical participants who readily gathered components whenever needed, those with ASD tended to wait until the cobot completed its task on its side of the table before gathering components. This preference for personal space likely contributed to increased waiting times for the cobot.
- **Varied Facial Expressions**: Interestingly, the ASD group displayed a wider range of facial expressions in response to the cobot's actions compared to the neurotypical group.
- **More other manifestations**: The ASD group showed a variety of "other manifestations", often involving their body (greeting, frowning, jumping, swaying).
- **Variability in behaviors**: Overall, there was greater variability in behaviors within the ASD group. The neurotypical group, on the other hand, exhibited more homogeneous behaviors.
- **Less talking with others**: Participants with ASD generally talked less with others in the room compared to the neurotypical group.

6.5.4 Behavioral Profiles of ASD Participants

This section summarizes the behavioral patterns of participants with ASD, compiled from persona cards created based on the live notes taken during three work shifts.

Task

While interacting with the cobot participants with ASD encountered some challenges. These included delays caused by missing cobot components that needed to be refilled, seeking technical assistance for minor issues they could potentially handle themselves, and cobot pauses caused by participant errors.

On the other hand, some participants also displayed positive behaviors that enhanced task performance. These strengths included the ability to talk and work simultaneously without getting distracted, a good understanding of the system's functions (such as knowing how to respond when the cobot has trouble finding a component or using the foot pedals correctly), and autonomy in managing their tasks (for example, rearranging their workstation).

Work Organization

The observational notes revealed varying levels of ability in terms of how ASD participants organized their work. In terms of keeping the cobot table stocked, some participants were able to refill the table with components while the cobot was performing its tasks. However, others had difficulty managing this activity. Similarly, not all participants consistently had their sub-assemblies prepared when the cobot arrived for collaboration. This lack of readiness sometimes led to delays in the overall assembly process.

Planning and organization skills also varied among participants. Some participants were adept at organizing multiple sub-assemblies in advance. This allowed them to stay ahead of the assembly and take advantage of downtime caused by cobot pauses to arrange components on the desk. However, this planning ability was not uniformly observed across all participants.

Break management also differed among participants. Ideally, taking breaks is important to avoid fatigue and maintain focus. While some participants were able to take breaks on their own initiative, others were so absorbed in the task that they required reminders from the experimenters. To manage physical fatigue, some participants used chairs to sit down.

An interesting behavioral pattern was observed relating to how participants handled the end of their shift. Many participants exhibited an aversion to leaving things unfinished. This sometimes led them to prioritize completing the current component box (which typically contained components for five gearboxes) or finishing all the components on the table.

Say - Quotes

This category focused on analyzing verbal statements from participants with ASD. Their quotes were grouped by common themes and presented in Table 6.9 for categories not already covered in other sections.

Interestingly, the quotes indicate a tendency among ASD participants to anthropomorphize the cobot. This observation is consistent with the findings of Atherton and Cross [2018] suggesting a tendency towards anthropomorphism and stronger empathic skills when interacting with non-humans in ASD populations.

Act - Recurrent Behaviors

The ASD participants exhibited some recurring behaviors during the experimental sessions. Some participants engaged in personal activities during the interaction, such as checking their cell phone or listening to music with headphones. Physical cues provided insights into their physical state, with behaviors like leaning on the table, sitting, stretching, puffing, or yawning potentially indicating exertion or fatigue.

Positive expressions were also observed, including giggling, humming, or keeping time with their foot, which could suggest enjoyment or engagement with the task. Participants also engaged in verbal behavior, such as chatting with others or even talking to themselves. Finally, some repetitive hand movements (for example, snapping fingers) were noted.

Feel - Emotion Expression

There were broadly four emotions that were identified from the notes: nervousness, boredom/tiredness, happiness, and fear.

Anthropomorphism
“Does the robot have a name?"
“Come on FANUC come on!" (referring to the cobot one more time looking for the parts it cannot find)
“I am sorry that you are waiting" (referring to the cobot) “How empathetic you are"
“Come on, there are three beautiful little pieces... Now I’m going to move it for you sweetie"
Attention to Details
“This piece is defective" (they realize that one piece is slightly different from the others)
“Maybe that’s why he’s having a hard time catching it" (they notice that one component is darker in color)
“I realized that by putting the smaller rings near the edge the cobot was not taking them"
Control/Feedback
“I need to calculate how long it takes me to do an assembly so that I will not leave any pieces for my colleague"
“I made half of this box, at the end of the week can you tell me how many pieces I made on average?"
“Will you count the assemblies or shall I count them?"
Opinion on Task
“While doing this work, those who are not Aspergers become so"
“It is relaxing for me to do this stuff, I don’t think while I am working, I have less pressure"

Table 6.9: Some quotes collected from the ASD participants during the live note-taking

Several factors triggered nervousness in participants, including long cobot pauses leading to inactivity, work disruptions from phone notifications, repeated cobot failures in detecting components, and the inability to finish the shift by completing the opened component box or all the components on the table.

Boredom or tiredness manifested through behaviors like puffing, slumping on the table, yawning, or sighing. Happiness was expressed through smiles, enjoying music, dancing, giggling, and humming. A sense of safety and comfort near the cobot was also observed. In some cases, participants exhibited fear by jumping when the cobot approached them.

Inference

The analysis of persona cards revealed diverse behavioral patterns in participants with ASD interacting with a cobot. Many of these patterns were captured by the observational grid. Therefore, this analysis further validates the findings from the observational grid regarding the ASD participants.

6.6 Quantitative Analysis

Building on the qualitative insights from Section 6.5, this section delves into quantitative comparisons of behavioral and performance patterns between neurotypical and ASD groups. By leveraging the NOVA annotations, the time spent by each group on various actions were assessed and compared.

6.6.1 Cobot Wait Duration

One of the key quantitative measures that differed significantly between the two groups was the cobot's waiting time. As mentioned in the qualitative analysis (Section 6.5), participants with ASD generally displayed a lower sense of urgency when responding to the cobot's waiting action.

To quantify this observation, the average waiting time experienced by the cobot was calculated across all video sessions for each group. The results supported the qualitative observations. On average, the cobot waited for neurotypical participants for only 20.7 seconds per video. In contrast, participants with ASD caused the cobot to wait for an average of nearly triple that duration, at 59.96 seconds per video. Figure 6.8 provides boxplots comparing the cobot's waiting time across all annotated videos for both groups. This visual representation helps in understanding the distribution of the cobot's waiting times within each group.

Statistical tests were conducted to determine if this difference was statistically significant. First, the mean cobot waiting durations for all participants were visualized using Q-Q plots. This analysis revealed that the data did not follow a normal distribution, violating an assumption for the commonly used independent samples t-test. Given the non-normal distribution of the data, the Mann-Whitney U test, a non-parametric alternative to the t-test, was chosen. This test confirmed a statistically significant difference in the cobot's waiting time between the two groups ($U = 11.0$, $p = 0.016$). This finding quantitatively supports

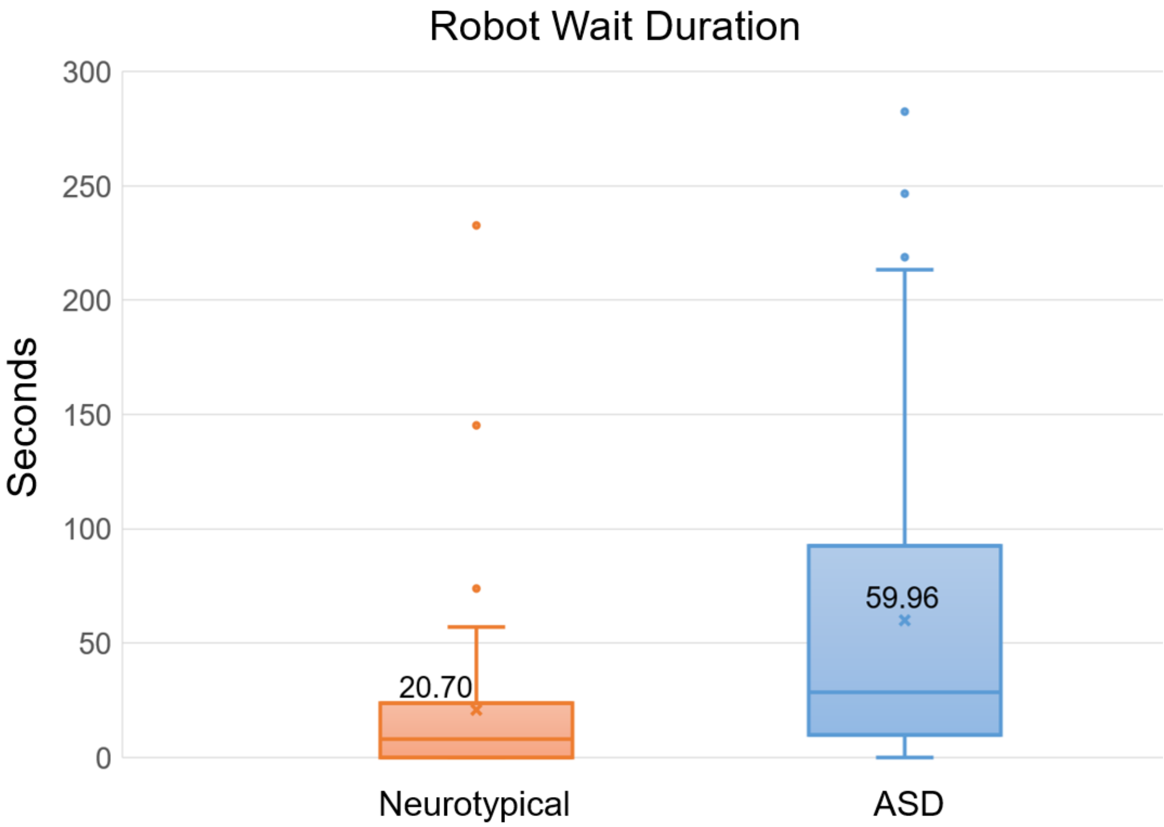


Figure 6.8: Box plot distributions of cobot wait durations for neurotypical and ASD groups. The x symbol in each distribution denotes the respective mean point.

the qualitative observation of a potentially lower sense of urgency among participants with ASD in attending to the cobot.

6.6.2 Gaze Duration and Gaze Continuity

The qualitative analysis suggested potential differences in how the two groups gazed at the cobot. This part of the analysis investigates these observations quantitatively, focusing on gaze patterns.

The analysis confirmed a significant difference in the amount of time participants spent looking at the cobot (Figure 6.9). On average, neurotypical participants gazed at the cobot for 52.02 seconds per video, whereas participants with ASD spent only 28.07 seconds per video looking at the cobot. This indicates that neurotypical participants exhibited nearly double the visual attention to the cobot compared to the ASD group.

Beyond gaze duration, the analysis also considered the duration of uninterrupted gaze contact with the cobot. Here, neurotypical participants again displayed a distinct pattern. They tended to have longer periods of continuous gaze towards the cobot, while participants with ASD exhibited shorter gaze durations and looked away more frequently. To quantify this disparity, the maximum duration of uninterrupted gaze contact was computed for each participant across all sessions. The average maximum gaze contact for participants with ASD was 7.93 seconds per video, compared to 12.49 seconds per video

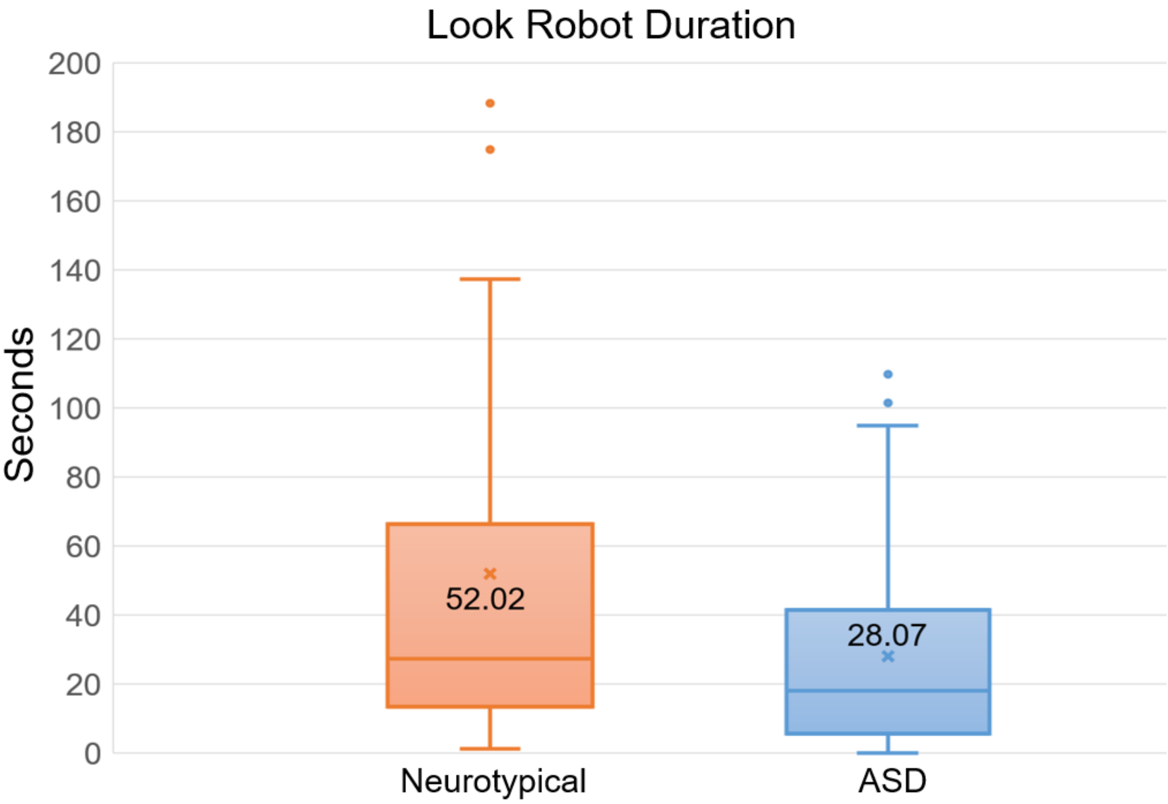


Figure 6.9: Box plots of durations of look robot action for neurotypical and ASD groups. The x symbol in each distribution denotes the respective mean point.

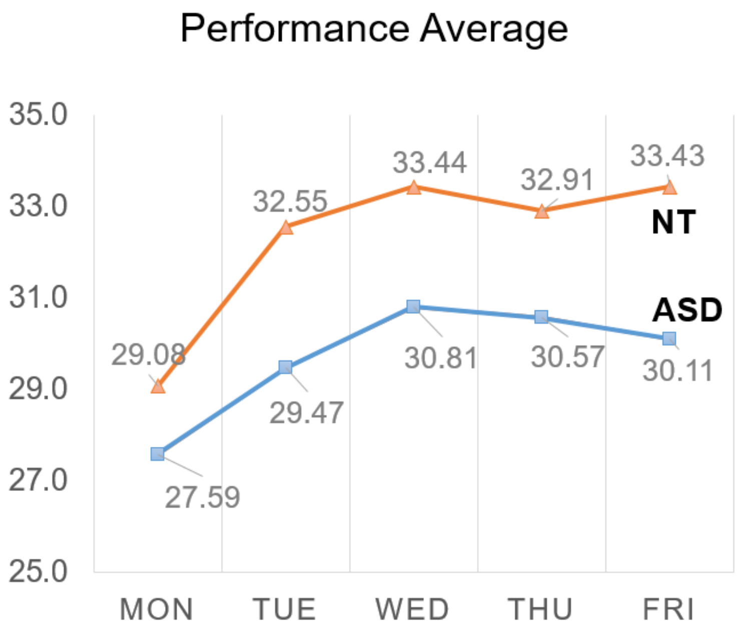
for neurotypical participants. These findings regarding gaze duration align with previous research by Damm et al. [2013] who observed a significant decrease in sustained gaze towards social robots among individuals with ASD over time.

Similar to the cobot's waiting time analysis, Q-Q plots were generated to assess the normality of the data for both look-at-cobot duration and maximum gaze contact duration. Neither measure followed a normal distribution, necessitating the use of non-parametric tests. The Mann-Whitney U test revealed a statistically significant difference in look-at-cobot duration between the groups (U = 15.0, p = 0.042). However, the maximum gaze contact duration yielded only a trend-level significance or p-value less than 0.1 (U = 19.0, p = 0.095). This suggests that a larger sample size might be necessary to detect potentially smaller effects in maximum gaze contact duration.

6.6.3 Performance

This section examines the performance data collected throughout the week-long experiment, revealing additional distinctions between the neurotypical and ASD groups.

The number of hourly assemblies for the neurotypical group (see Figure 6.10) exhibits a clear upward trend in performance for all participants. This is reflected in a 15% increase in the average performance index over the week (from 29.08 assemblies/hour on Monday to 33.43 assemblies/hour on Friday). Additionally, they seem to converge towards a common performance level. The daily standard deviations based on individual performance scores



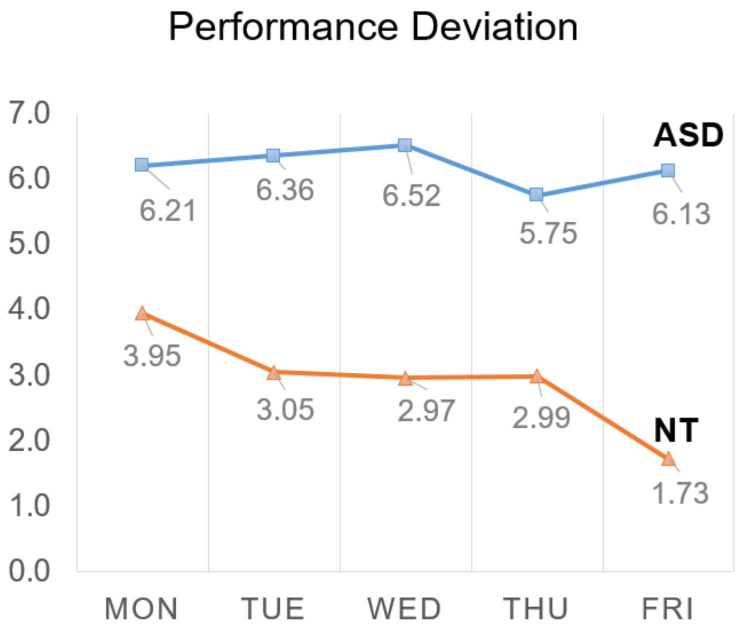


Figure 6.10: The plots represent the average (top) and standard deviation (bottom) of performance data (assemblies/hour) for neurotypical and ASD groups.

steadily decrease from 3.95 to 1.73, suggesting a collective improvement and convergence towards a similar level of performance by the end of the week.

The hourly assemblies of the ASD group (see Figure 6.10) also show a moderately increasing performance trend over the week (around 9% increase, from 27.59 to 30.11 assemblies/hour). However, unlike the neurotypical group, individual performance trends within the ASD group are more spread out, with no apparent convergence or divergence during the experiment. Daily standard deviations for the ASD group also remain relatively stable, ranging between 5.75 and 6.52, indicating less consistency in performance levels compared to the neurotypical group.

Interestingly, both the best and worst performers across all participants belonged to the ASD group. As seen from Figure 6.11, the performance range for the neurotypical group spanned between 24.57 and 38.75 assemblies/hour, while the ASD group exhibited a wider range of 19.50 to 41.74 assemblies/hour. This wider range aligns with the qualitative observations of greater behavioral variability within the ASD group compared to the more homogeneous neurotypical group.

To formally compare the two groups, the normality of the performance data was verified using Q-Q plots and Shapiro-Wilk tests (NT: p = 0.490, ASD: p = 0.094). Since the data for both groups appeared to be normally distributed, an ANCOVA test was conducted to analyze the influence of time and group membership on the collected performance indexes. This test confirmed a statistically significant difference between the two groups (F = 4.85, p = 0.010).

6.6.4 Impact of Downtime

To explore the potential relationship between downtime and performance, Figure 6.12 presents daily downtime trends for each participant. The plot on the left side shows data for the neurotypical group, while the plot on the right side shows data for the ASD group.

In this case, downtime includes both participant-initiated breaks and unexpected system pauses. Downtime could influence the level of fatigue, which in turn may affect the performance. However, examining the individual trends for both performance and downtime across participants reveals no clear correlation. For example, participant 3011004 (chosen due to their variable downtime) experienced a considerable increase in downtime followed by improved performance. Then, this participant had a substantial decrease in downtime followed by another performance increase. This lack of correlation between downtime and performance holds true for most participants, and similar conclusions can be drawn when examining their data. These observations suggest that the duration of downtime may not be a major factor influencing performance.

6.7 Insights

The qualitative and quantitative analyses complemented each other, reinforcing the findings. The following three behavioral patterns stand out as they emerged consistently in qualitative and quantitative measures.

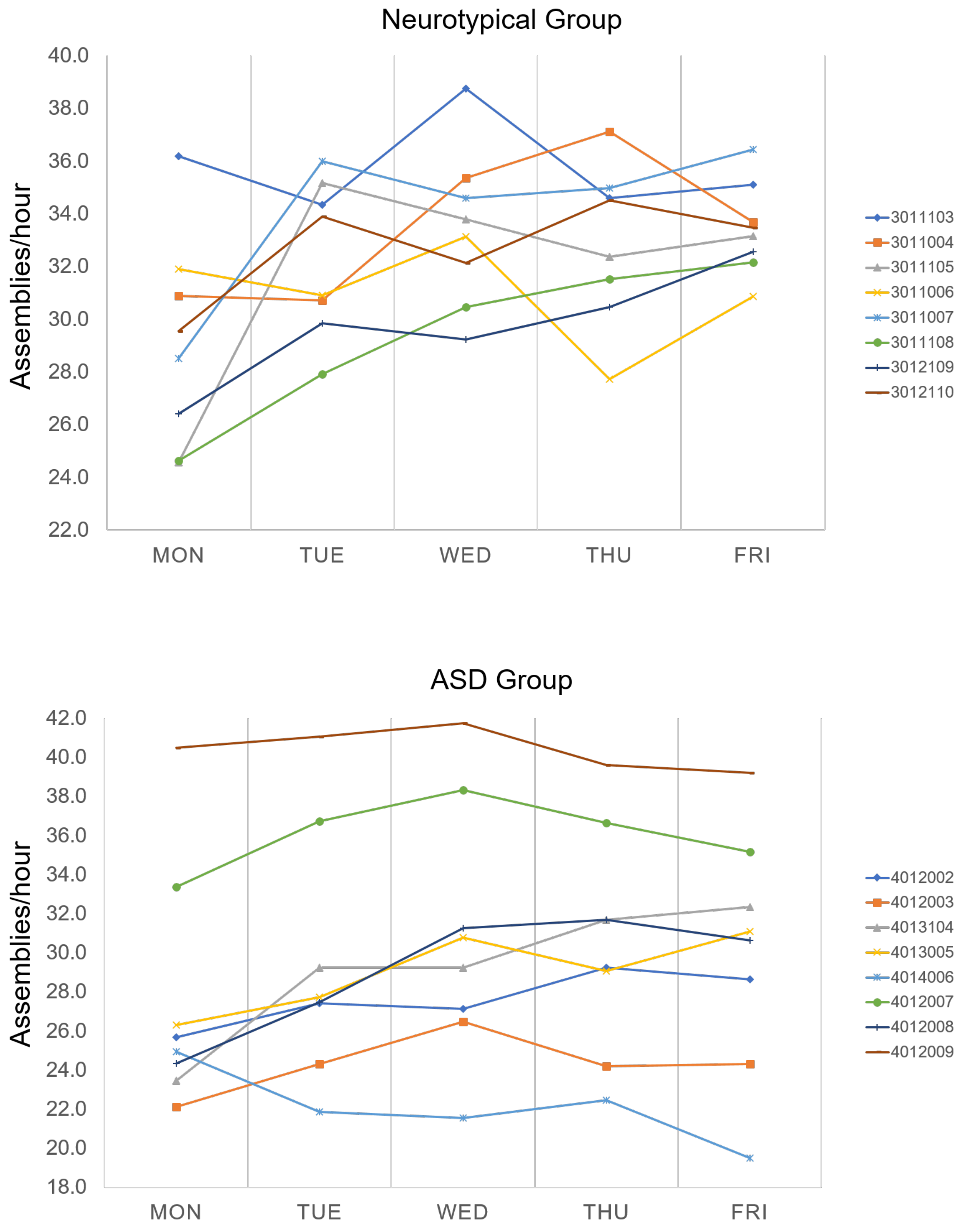


Figure 6.11: The plots visualize the performance data (assemblies/hour) for both neurotypical (top) and ASD (bottom) groups. Each line in the plot represents the assembly rate for a participant during the week.

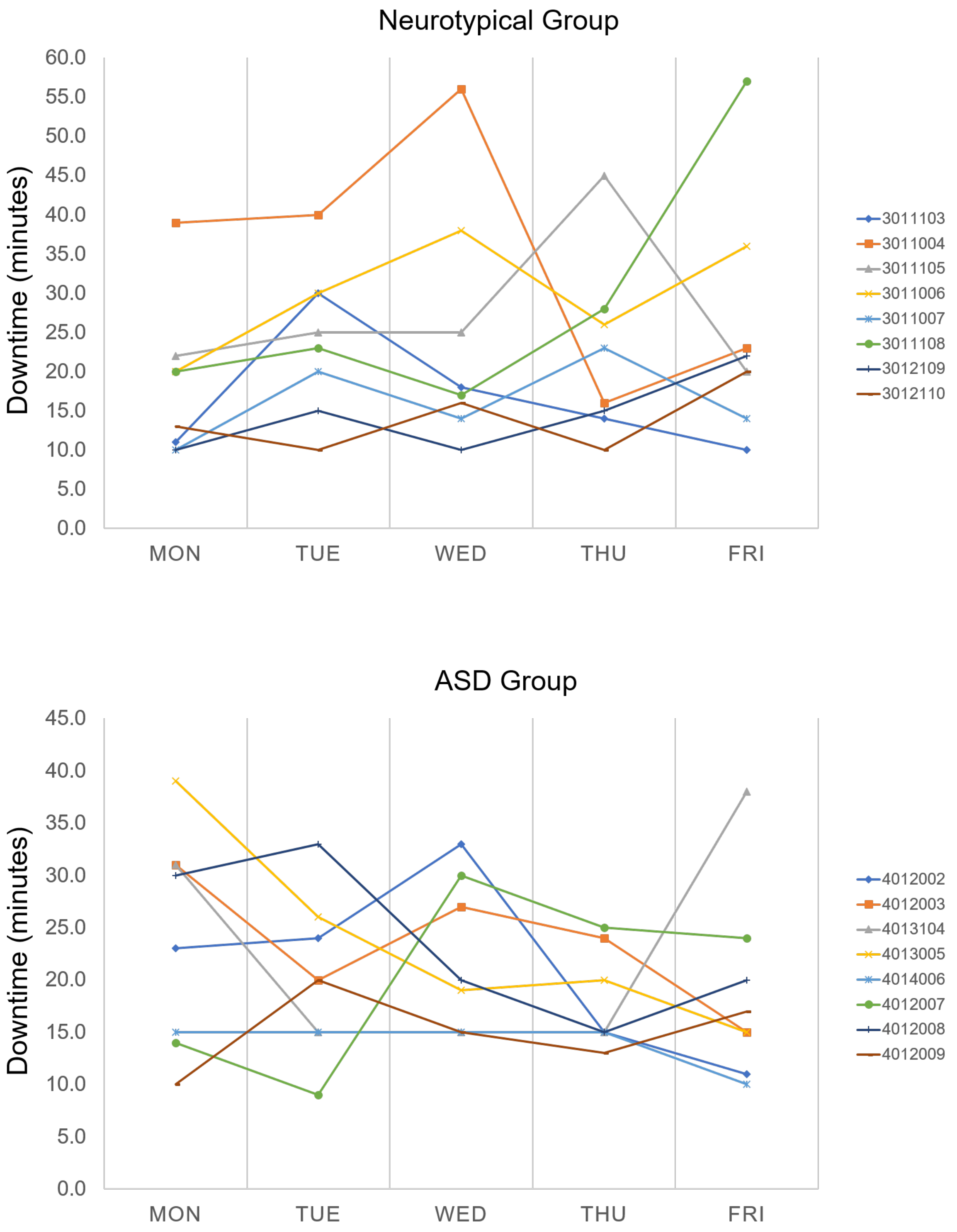


Figure 6.12: The plots visualize the downtime durations for both neurotypical (top) and ASD (bottom) groups. Each line in the plot represents the downtimes for a participant during the week.

- **Tiredness/Boredom**: Both neurotypical and ASD participants exhibited cues expressing tiredness or boredom. This state was probably caused by the disparity in the production rates of the participants and the cobot. The cobot was slower than the participants in assembling the individual sub-assembly, which led to participants waiting for a considerable amount of time. As the participants got more familiar with the task over the week, the manifestations of tiredness and boredom also increased.

- **Cobot Priority**: Participants with ASD exhibited a lower urgency in responding to the cobot's waiting action, often completing other tasks before attending to the cobot. This behavior led to longer waiting times for the cobot. This observation aligns with the findings of Murin et al. [2016], who observed task prioritization difficulties in ASD individuals. Conversely, neurotypical participants prioritized the cobot and the joint activity, resulting in minimal cobot waiting times. This difference in cobot prioritization likely impacted assembly performance, with the ASD group completing fewer assemblies on average.

- **Gaze Behaviors**: Both groups displayed a tendency to look at the robot, but the duration of gaze differed. Zhang et al. [2017] suggest that gaze information can enhance synchrony and communication in human-human collaboration. However, in this study, the intention behind the participant's gaze could not be discerned from the current data. Gaze could signal task completion to the cobot, facilitating synchronization. It could also indicate that the participants were monitoring the cobot's actions.

- **Assembly Routines and Performance**: Both groups generally improved their assembly performance over the week, with a steeper improvement rate for the neurotypical group. This suggests a learning curve for both groups initially, followed by optimization of work patterns (e.g., multitasking) by the neurotypical group in the latter days to achieve a better performance. This is further supported by the convergence of the neurotypical group towards a common performance index, reflecting a saturation point based on the task setup. Conversely, the ASD participants more or less maintained their working patterns, limiting their performance to the inherent efficiency of their initially chosen strategies.

6.8 Reflections and Remarks

This chapter presented an exploratory study that identified behavioral patterns in neurotypical and ASD participants collaborating with a cobot. The study analyzed the data collected from eight individuals from each group during a week-long experimental session. Some key differences emerged between the two groups including cobot prioritization, gaze patterns, and multitasking. Notably, while these findings align with existing research on how ASD individuals navigate social settings, it is interesting to observe similar patterns emerging in a context that is not overtly social, lacking human or humanoid interaction.

The study outcomes hold significant implications for Industry 5.0. A specific ASD participant outperforming their neurotypical counterparts highlights the immense potential for inclusivity within Industry 5.0 workplaces. The findings also hint that solutions designed for the neurotypical population may not effectively meet the needs of individuals with ASD. A personalized approach that caters to individual traits and preferences is crucial, particularly in designing adaptive cobot behaviors and balancing task loads.

It is important to note that the research design did not explicitly elicit responses to specific scenarios, such as cobot mistakes or handling stressful situations. These situations can significantly impact participants' responses and behaviors, and warrant further investigation.

Several key findings of this chapter form the foundation for the research presented in the subsequent chapters. The production rate of the collaborating partners (cobot and operator) is a crucial aspect to be considered in shaping the collaboration experience. For example, long periods of waiting could result in the experience of boredom and tiredness. This aspect is exploited in Chapter 7 to study the states like boredom, anxiety, and flow. Another interesting aspect to consider is gaze behaviors in HRC settings and how they can be used to facilitate collaboration. Chapter 8 delves deeper into this topic, examining the potential of cobot adaptations based on gaze cues.

Moreover, a subset of the annotated video data used in the quantitative analysis of this chapter was leveraged for an additional purpose. Chapter 3 presents an assessment of the applicability of an attention recognition model in industrial HRC settings using this dataset.

Chapter 7

Flow in Industrial HRC

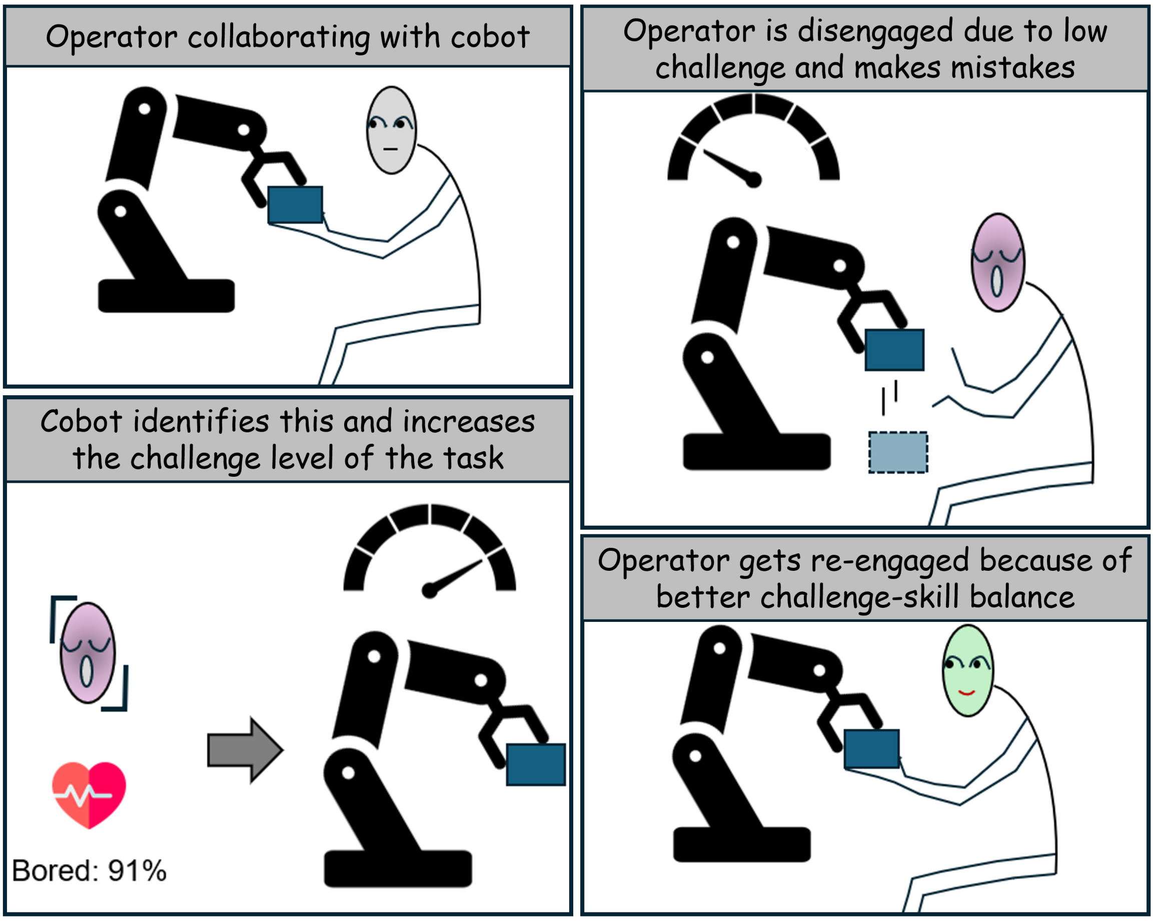

Operator collaborating with cobot
Operator is disengaged due to low challenge and makes mistakes
Cobot identifies this and increases the challenge level of the task
Bored: 91%
Operator gets re-engaged because of better challenge-skill balance

Figure 7.1: A comic strip illustration of how perceived challenge detection can improve the collaboration experience in an industrial scenario. In this situation, the operator experiences boredom due to the low challenge level of the current task. The cobot detects the low perceived challenge and modifies the challenge level of the task to facilitate the experience of flow.

7.1 Overview

This chapter explores flow, a state of optimal experience characterized by high engagement, control, and immersion [Csikszentmihalyi, 1975, 2020; Nakamura and Csikszentmihalyi, 2002], in industrial Human-Robot Collaboration (HRC) settings. As mention in Chapter 2, the balance between challenge level of the task and skill of the individual is a necessary condition for experiencing flow state. Flow has been extensively studied in various domains such as sports [Stamatelopoulou et al., 2018], education [dos Santos et al., 2018; Pearce, 2005], and gaming [Nah et al., 2014], but its application in industrial contexts remains relatively unexplored. Flow is particularly relevant to industrial HRC due to its potential to enhance worker well-being, performance, and job satisfaction [Csikszentmihalyi, 2020; Maeran and Cangiano, 2013; Fullagar et al., 2018; Christandl et al., 2018; Peifer et al., 2020; Peifer and Wolters, 2021]. When workers experience flow in collaborative tasks, they are more likely to be engaged, focused, and make fewer errors. This can lead to improved productivity, reduced stress, and a more positive work environment. Figure 7.1 depicts a hypothetical industrial HRC scenario where the cobot adapts task difficulty to maximize flow among cobot workers.

This chapter investigates the perceived challenge levels and their corresponding emotional and physiological responses in an industrial assembly task involving a cobot. Facial emotion estimation (valence and arousal) and heart rate variability are explored as potential indicators of perceived challenge levels. A predictive model was developed to estimate perceived challenge levels based on these indicators. The findings of this chapter could be used to develop adaptive HRC systems that automatically adjust task difficulty to match the perceived challenge levels of individual workers, fostering optimal worker experiences and enhancing productivity.

The contents of this chapter including the setup, analyses, and models have been presented previously in:

* P. Prajod, M. Lavit Nicora, M. Mondellini, M. M. Falerni, R. Vertechy, M. Malosio, and E. André. Flow in human-robot collaboration—Multimodal analysis and perceived challenge detection in industrial scenarios. *Frontiers in Robotics and AI*, 11:1393795, 2024a

  [*I contributed significantly to the study design and formulation of the hypotheses. I also performed data processing, feature engineering, and development of the machine learning models. Furthermore, I conducted the analysis and derived insights.*]

* M. Mondellini, M. L. Nicora, P. Prajod, E. André, R. Vertechy, A. Antonietti, and M. Malosio. Exploring the dynamics between cobot's production rhythm, locus of control and emotional state in a collaborative assembly scenario. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2024

  [*I contributed significantly to the study design and formulation of the hypotheses. Additionally, I contributed to selection of analysis tools.*]

* F. Nunnari, M. L. Nicora, P. Prajod, S. Beyrodt, L. Chehayeb, E. André, P. Gebhard, M. Malosio, and D. Tsovaltzi. Understanding and mapping pleasure, arousal and

dominance social signals to robot-avatar behavior. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2023

[*I trained the emotion recognition model and developed the data processing and detection pipeline.*]

* S. Beyrodt, M. L. Nicora, F. Nunnari, L. Chehayeb, P. Prajod, T. Schneeberger, E. André, M. Malosio, P. Gebhard, and D. Tsovaltzi. Socially interactive agents as cobot avatars: Developing a model to support flow experiences and well-being in the workplace. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2023

  [*I contributed to study design and selection of analysis tools.*]

7.2 Previous Works

7.2.1 Flow and Emotions

Flow has been linked to positive affective states such as enjoyment and engagement. Conversely, when the perceived challenge and individual skill level are mismatched, negative emotions such as anxiety and boredom can arise. Previous research has explored the emotional responses associated with flow, mapping the flow state in terms of valence, arousal, and dominance. Valence reflects the pleasantness or unpleasantness of an emotional state, arousal indicates the level of activation linked with an emotional response, and dominance refers to perceived control over the emotional experience. The following literature, identified through a Scopus database search[1], offers insights into the connection between flow and various dimensions of emotion.

To investigate the connection between flow and emotional dimensions of valence and arousal, Kivikangas [2006] utilized a digital game to induce flow experiences. Facial muscle activity (measured through facial Electromyogram; facial EMG for short) and Electrodermal activity (EDA) were measured as proxies for valence and arousal, respectively. Specifically, three facial muscles were monitored: the corrugator supercilii (CS) associated with frowning (negative valence), the zygomaticus major (ZM) associated with smiling (positive valence), and the orbicularis oculi (OO) associated with widened eyes (positive valence). As expected, flow experiences exhibited a negative correlation with CS activity, suggesting a decrease in negative emotions during flow. However, EDA, an indicator of arousal, surprisingly revealed no link to flow. Additionally, results for ZM and OO muscles representing positive valence remained inconclusive, possibly due to limitations in data measurements.

Building upon the previous work of Kivikangas, Nacke and Lindley [2008] also employed facial EMG and EDA to measure participants' emotional responses while playing games. They induced flow and boredom states from the three-channel model. They observed that increasing the game challenge did not evoke frustration, but instead induced

[1] `https://www.scopus.com/`, Query: (TITLE-ABS-KEY ((emotion OR affect) AND (valence OR arousal OR dominance)) AND TITLE (flow))

feelings of enjoyment due to engaging gameplay. Interestingly, they observed contrasting results compared to Kivikangas, revealing significant differences in EDA, ZM, and OO muscle activations. Their findings suggest that flow manifests as a state of positive valence with high arousal. However, the authors acknowledged that their study was limited to a specific demographic of male experienced gamers, potentially limiting the generalizability of the findings.

Gilroy et al. [2009] explored the relationship between flow and emotion dimensions by utilizing an augmented reality interactive art installation. Their focus was to map the flow experience onto the dimensions of arousal and dominance. They proposed that both anxiety and boredom represent low-dominance states, with anxiety characterized by high arousal and boredom by low arousal. In contrast, the flow experience was characterized as a high-dominance, high-arousal state. While a preliminary study with limited participants supported this proposed mapping, further investigation is needed to validate this model.

To explore whether social networking induces flow, Mauri et al. [2011] explored physiological responses elicited by Facebook usage compared to relaxation induced by nature slideshows, and stress induced by the Stroop test. Mapped onto the valence-arousal model, they hypothesized relaxation as a positive-valence, low-arousal state, stress as a negative-valence, high-arousal state, and Facebook usage resembling a flow state characterized by high arousal and positive valence. To measure participants' physiological responses, the study gathered various signals including EDA, facial EMG, Blood Volume Pulse (BVP), and pupil dilation. Focusing on the CS muscle activation as an indicator of valence and EDA as an indicator of arousal, the authors plotted their findings onto the valence-arousal space. Their data aligned with the hypothesized mapping, suggesting that Facebook usage indeed occupied a region close to the high-arousal, positive-valence area associated with flow experiences.

While the previous research suggested a positive relationship between flow and arousal, Peifer et al. [2014] demonstrated an inverted U-shaped relationship. Their argument centered around the limitations of typical lab studies, which often utilized game-based tasks that might not be perceived as personally relevant or stressful enough to induce very high arousal states. To address this limitation, they employed a two-step approach. Participants were first subjected to a socially stressful situation. They then engaged in a computer task designed to induce boredom, anxiety, and flow states. They analyzed the low-frequency (LF) and high-frequency (HF) components of Heart Rate Variability (HRV) as indicators of arousal and relaxation. Their findings confirmed an inverted U-shaped relationship, with flow occupying a state of moderate arousal between the boredom and anxiety extremes.

Figure 7.2 depicts the positioning of boredom, anxiety, and flow states within the valence-arousal-dominance scales, as derived from the literature. Flow is characterized by high dominance and positive valence, distinguishing it from boredom and anxiety, which are associated with low dominance and negative valence. The distinction between boredom and anxiety is further refined by considering the arousal dimension. This visualization, supported by aforementioned studies, underscores the effectiveness of using two-dimensional models like valence-arousal or arousal-dominance to differentiate between boredom, anxiety, and flow experiences.

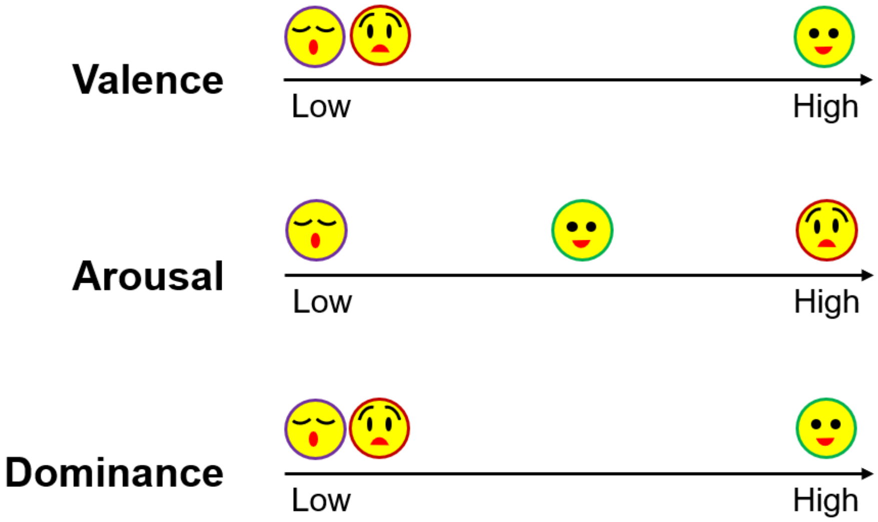


Figure 7.2: A visualization of how flow, boredom, and anxiety map onto the dimensions of valence, arousal, and dominance, based on existing research. The copyright remains with the authors [Prajod et al., 2024a].

7.2.2 Flow and HRV features

Numerous studies Knierim et al. [2018]; Khoshnoud et al. [2020] have explored the flow experience using different physiological signals, such as HRV, EDA, and respiration. This interest stems from the relationship between flow and arousal (see Figure 7.2), making physiological responses promising modalities for investigation. This investigation is particularly crucial for the automatic detection of flow.

Although EDA appears as a prevalent measure of arousal in the works discussed in Section 7.2.1, recent research investigates HRV as an alternative. This is plausibly due to its superior discriminating power for flow detection Knierim et al. [2018]. Additionally, findings in the next section (Section 7.2.3) suggest that incorporating EDA doesn't necessarily enhance flow detection performance. Therefore, this chapter primarily focuses on the relationship between flow and HRV.

Flow is linked to the activity of the autonomic nervous system (ANS), which regulates various bodily functions. Sympathetic activation of the ANS results in arousal, while parasympathetic activation leads to relaxation. Notably, sympathetic activation increases heart rate, while parasympathetic activation is associated with a decrease in heart rate [Pham et al., 2021]. Both sympathetic activation and parasympathetic activation of the ANS have been associated with the different experiences of the three-channel Flow model [Knierim et al., 2018]. As HRV features effectively capture both types of ANS activation, they are well-suited for analyzing experiences within the three-channel flow model. In particular, the flow state is often characterized by increased heart rate and reduced inter-beat intervals or mean HRV Tian et al. [2017]. This section discusses relevant literature identified through a Scopus database search[2], specifically focusing on works that investigated the relationship between flow and HRV features.

[2] `https://www.scopus.com/`, Query: (TITLE-ABS-KEY ((experience OR state) AND ("heart rate variability" OR hrv)) AND TITLE (flow))

As one of the early investigations into physiological responses to flow, de Manzano et al. [2010] examined various physiological signals in professional pianists, including HRV (derived from BVP), facial EMG, and respiration. They analyzed mean HRV and power spectrum features (LF/HF ratio, total power), and found significant correlations between flow intensity and all three features. Notably, mean HRV showed an inverse relationship with flow, suggesting that HRV decreased with an increase in flow. Conversely, frequency features showed a positive correlation with flow. These findings were interpreted as evidence supporting a relative increase in sympathetic activation during flow, indicating heightened arousal.

Expanding on the research conducted by de Manzano et al., Jha et al. [2022] investigated the HRV responses of pianists before, during, and after a performance. Their analysis focused on the low- and high-frequency components of the HRV signal derived from Electrocardiogram (ECG) data. They found that lower sympathetic activity before the performance, as indicated by the LF/HF ratio and LF, predicted peak flow, suggesting a relaxed pre-performance state is crucial for experiencing flow during the performance. Additionally, similar to de Manzano et al., they also found a trend towards a positive correlation between LF/HF ratio and flow, indicating a link between sympathetic activity or arousal and the flow state.

Investigating the relationship between mental load and physiological responses, Keller et al. [2011] studied how HRV features varied during underload, overload, and fit conditions of a computerized quiz game. They collected ECG data and measured cortisol levels as markers of stress. As expected, the underload condition with the lowest mental load showed the highest HRV. Although the fit condition showed decreased HRV compared to the underload condition, this decrease could represent either high engagement or mental strain. To discern between the two possibilities, cortisol levels were analyzed. Interestingly, despite participants reporting subjective flow experiences in the fit condition, their cortisol levels were comparable to those observed in the overload condition.

Similarly, Peifer et al. [2014] studied the flow experience under stressful conditions utilizing a computerized game. They analyzed low-frequency (LF) and high-frequency (HF) components of HRV as indicators of sympathetic and parasympathetic activity, respectively. The LF component showed an inverted U-shaped relationship with flow experience, suggesting a moderate arousal for flow. Additionally, HF showed a positive linear association with flow, indicating increased relaxation. These findings further corroborate the idea that both sympathetic and parasympathetic activities contribute to flow experiences.

Expanding on previous research, several studies have utilized games featuring varying challenge levels (underload, overload, and fit) to explore the physiological responses associated with the flow experience. They measured multiple physiological signals and correlated them to the experience of flow. Given the similarity in the outcomes of HRV analyses across these studies, they are collectively summarized below.

While Harmat et al. [2015] observed that flow was associated with low LF, their results lacked statistical significance. However other studies like Bian et al. [2016], Tian et al. [2017], and de Sampaio Barros et al. [2018] found a linear relationship between difficulty level and both heart rate (lowest for underload, highest for overload) and mean HRV (highest for underload, lowest for overload), indicating arousal due to sympathetic activation.

Additionally, Bian et al. found an inverted U relationship between flow and both LF and HF components of HRV.

Following a similar approach to video game studies, Tozman et al. [2015] examined the link between HRV and flow states using a virtual driving simulator with varying difficulty levels (boredom, anxiety, and optimal challenge). In addition to the driving challenge, they also incorporated a social evaluation stressor into the anxiety condition. They collected ECG data of the participants and analyzed the LF and HF components of the derived HRV signal. They found higher task difficulty led to decreased LF, with the highest levels observed during boredom, moderate during the fit condition, and lowest during anxiety. A similar negative linear relationship was observed between flow and HF spectrum components. While the decrease in parasympathetic activity (lower HF, lower relaxation) with higher difficulty aligns with expectations, the decrease in sympathetic activity (lower LF, lower arousal) contradicts some existing literature that associates flow with arousal. The authors note that LF can be interpreted as a measure of baroreflex activity, a feedback mechanism through which the body adjusts heart rate in response to sudden changes in blood pressure, both decreasing it when pressure rises and vice versa.

Expanding beyond lab settings, Gaggioli et al. [2013] investigated flow experiences within the daily lives of university students. Leveraging a week-long study, they monitored students' ECG data and revealed connections between flow and specific HRV features. Notably, they found positive correlations between flow experiences and both heart rate and LF/HF ratio. These increases suggest a relative increase in sympathetic activation, aligning with lab studies suggesting arousal during flow. However, their study lacked analysis of the nature of the activities participants engaged in or the type of challenge (mental, physical, etc.).

Emulating a real-world scenario, Knierim et al. [2019] explored the experience of flow in knowledge tasks. They used arithmetic with three difficulty levels (boredom, flow, overload) and scientific writing as the knowledge tasks. They computed two ECG-based HRV features: RMSSD (lower value indicates higher stress) and HF. In the arithmetic task, the flow condition showed higher RMSSD compared to overload, indicating lower stress. They also observed a trend-level decrease in HF with the increase in challenge, consistent with the literature. However, the writing task did not exhibit these patterns. Interestingly, both RMSSD and HF were consistently lower in the writing task compared to the arithmetic task, suggesting a potentially higher arousal state. This study highlights that even similar knowledge tasks can evoke different flow experiences and physiological responses.

Most studies investigating flow primarily focus on mentally demanding tasks and often analyze HRV features such as heart rate, mean HRV, LF, and HF. Heart rate often correlates positively with task challenge, while mean HRV exhibits a negative linear relationship. The HF component is typically associated with parasympathetic activity and the others with sympathetic activation. However, findings regarding the relationship between flow experience and frequency components vary across studies, likely due to differences in study design [Knierim et al., 2019].

Paper	Scenario	Relating HRV features
de Manzano et al. [2010]	Piano	mean HRV, LF/HF, total power
Jha et al. [2022]	Piano	LF, HF, LF/HF
Keller et al. [2011]	Game	mean HRV
Peifer et al. [2014]	Game	LF, HF
Harmat et al. [2015]	Game	LF
Bian et al. [2016]	Game	HR, mean HRV, LF, HF
Tian et al. [2017]	Game	HR, mean HRV
de Sampaio Barros et al. [2018]	Game	HR, mean HRV
Tozman et al. [2015]	Driving simulator	LF, HF
Gaggioli et al. [2013]	Daily student activities	HR, LF/HF
Knierim et al. [2019]	Knowledge tasks	RMSSD, HF
Prajod et al. [2024a]*	**HRC**	13 features, delved into HR, mean HRV, LF, HF

*Table 7.1: An overview of works that linked flow experience and HRV features. The entry marked with * is expanded in the subsequent sections of this chapter.*

7.2.3 Flow Detection at Workplace

Some studies have introduced automatic flow detection models utilizing physiological signals within gaming contexts [Khoshnoud et al., 2020]. These models typically induce varying levels of difficulty or challenge in the game and classify the corresponding physiological data. Recently, some studies have extended this approach to activities within workplace settings. The following works were identified through a Scopus literature search[3] for studies that trained machine learning models for predicting flow at work.

For instance, Müller and Fritz [2015] investigated the feasibility of predicting whether programmers were “stuck" or adequately progressing during software development tasks. They considered a high-progress rating as an indicator of being in the flow state. They employed a multimodal approach, collecting various physiological signals including HRV (derived from BVP), Electroencephalogram (EEG), pupil features, and EDA. Utilizing these

[3] `https://www.scopus.com/`, Query: (TITLE-ABS-KEY ((“at work" OR worker OR workplace OR employee) AND (detect* OR recogni* OR classif*) AND (“deep learning" OR ml OR “machine learning" OR network)) AND TITLE (flow))

signals, they trained a machine learning classifier to distinguish between low-progress and high-progress instances, achieving an accuracy of 63.35% in leave-one-subject-out (LOSO) evaluations.

Similarly, Lee [2020] employed a multimodal approach, utilizing physiological signals like HRV (derived from BVP), EDA, and pupil diameter. The study, conducted in a controlled lab setting, involved researchers and graduate students engaging in diverse knowledge tasks such as editing spreadsheets, reading and summarizing text, and answering patent questions. Features extracted from the physiological signals were used to train various machine learning models for a variety of classification tasks including flow recognition and working state detection. Notably, the binary classifier for distinguishing flow from non-flow states achieved an AUC of 0.889 in LOSO evaluations.

Further investigating the applicability of flow detection models in a workplace, Rissler et al. [2020] introduced machine learning models to classify low-flow and high-flow instances. They relied solely on ECG-derived HRV features for training their models. They conducted two experiments: a controlled lab study with an invoice-matching task involving varying difficulty levels, and an in-the-field study with software developers performing their regular tasks. Their model achieved an accuracy of 68.5% in the lab setting and even higher (70.6%) in the real-world study, demonstrating the potential for automatic flow detection in actual workplace environments.

Further building upon the concept of in-the-field models, Di Lascio et al. [2021] explored the use of physiological signals (BVP, EDA) alongside contextual information to predict low-flow and high-flow instances. Their study recorded the physiological signals of university employees (professors, researchers, PhDs) during various daily activities. While individual modalities like HRV offered promising accuracy (67.46%), the most successful model achieved an accuracy of 70.93% by fusing raw BVP, EDA, and contextual information. However, despite the availability of additional modalities, the accuracy of their best model was comparable to the HRV-only model from Rissler et al..

The studies discussed thus far showcase the potential of flow detection in work settings. However, they primarily focus on mentally demanding tasks like knowledge work and software development and often involve specific participant groups like researchers or graduate students. These characteristics may not directly translate to the industrial scenario, where tasks and participant demographics differ significantly.

7.2.4 Research Gaps

Through the above literature review, the following research gaps are identified, primarily stemming from the lack of research on the experience of flow in the HRC context. The experiments described in this chapter were designed to address these gaps.

- **Facial emotion estimations as indicators of flow**: Previous research (Section 7.2.1) linked the three-channel flow model to the valence-arousal-dominance emotional dimensions, suggesting that flow states can be characterized by arousal combined with either valence or dominance. Moreover, some works have used facial EMG features as a measure of valence. Recent advancements in affective computing have enabled the estimation of emotional dimensions, especially valence and arousal, from facial

Paper	Demand	Physiological signals	Accuracy
Müller and Fritz [2015]	Mental	BVP, EEG, EDA, Pupil	63.35%
Lee [2020]	Mental	BVP, EDA, Pupil	–
Rissler et al. [2020]	Mental	ECG	70.6%
Di Lascio et al. [2021]	Mental	BVP, EDA, Context	70.93%
Prajod et al. [2024a]*	**Physical, Temporal**	ECG	70.7%

*Table 7.2: An overview of works on flow detection in the workplace. The entry marked with * is expanded in the subsequent sections of this chapter.*

images. However, although these methods are non-intrusive, they are seldom employed in flow detection. A notable exception was the work by Burns and Tulip [2017]), who analyzed facial expressions (valence and arousal) during a video game with varying challenge levels. Given the limited existing research, further investigation is necessary to assess the effectiveness of facial emotion estimation methods in identifying flow states, especially within HRC settings.

- **HRV features as indicators of flow**: Previous research (see Section 7.2.2) has extensively explored the relationship between HRV and the three-channel flow model. As highlighted by Table 7.1 existing research focuses on tasks where challenge is induced primarily through mental load. However, industrial HRC often involves tasks with significant physical demands and time pressure, raising questions about the applicability of these findings. Specifically, it remains unclear whether similar HRV responses are elicited when challenges arise from factors beyond mental demand.

- **Automatic flow recognition**: The literature survey presented in Section 7.2.3 reveals that there are few studies that developed machine learning models for flow detection in the workplace. Moreover, similar to HRV analysis, these studies also primarily focus on mentally demanding tasks (see Table 7.2). However, the task challenges in HRC extend far beyond mental load, encompassing factors such as cobot speed, production rate, and interaction dynamics. This gap is further compounded by the lack of understanding about which flow-related states are most critical to detect in HRC settings. Hence, there is a need to train flow recognition models specifically tailored for HRC scenarios, with the aim of adapting scenario-specific parameters to facilitate flow experiences.

7.3 Balancing Challenge and Skill

Repetitive and fixed procedures are common characteristics of assembly tasks in industrial settings. Workers typically gain proficiency in these tasks, leading to stable individual skill levels over time. This trend was observed in the analyses of the week-long study described in Chapter 6. Consequently, the perceived challenge level of the task becomes the primary determinant of the flow experience. Recognizing that flow emerges when perceived challenge aligns with skill level, the goal is to develop adaptive task systems that dynamically adjust challenge levels to foster flow among cobot workers.

Previous research in HRC [Kulić and Croft, 2007; Arai et al., 2010; Koppenborg et al., 2017; Kühnlenz et al., 2018; Gervasi et al., 2022; Zakeri et al., 2023] has studied the effects of robot movement speed and proximity on operators' stress and anxiety levels. However, these factors have not been shown to elicit boredom and flow. Insights from Chapter 6 revealed that production rate could potentially induce boredom and flow, in addition to anxiety. The procedure employed to evoke these states is detailed in the following subsections.

7.3.1 Setup and Assembly Task

The experiment takes place in a controlled laboratory setting designed to resemble an industrial collaborative robotic work cell. The setup and assembly tasks were largely similar to Chapter 6. However, slight adjustments were implemented to facilitate control over the cobot's production rate. Unlike Chapter 6 where the cobot assembled its own sub-assemblies alongside the participant, this study employed pre-assembled sub-assemblies readily available for the cobot to pick and deliver. This choice enabled manipulation of the cobot's production rate across various experimental conditions.

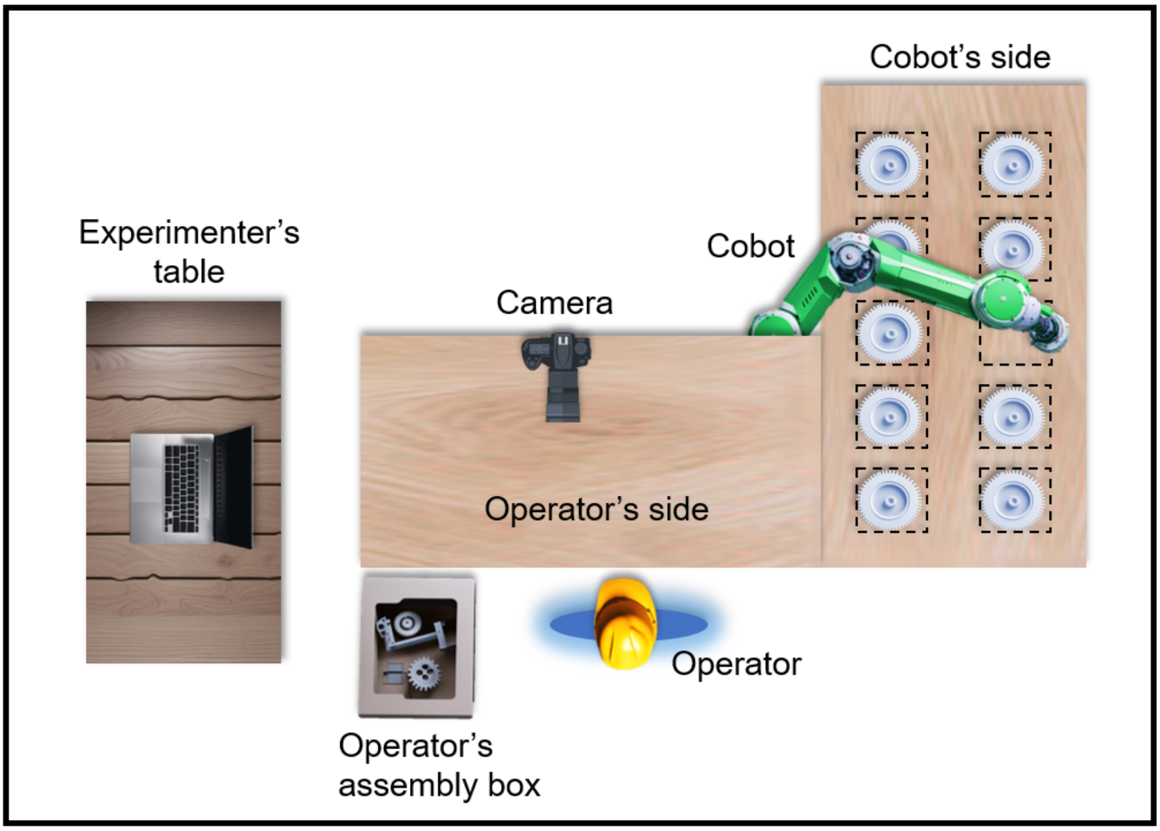
Task

The collaborative assembly task involves both the participant and the cobot working together to build a planetary gearbox (see Chapter 6). While the participant assembles half the components following specific instructions, the cobot delivers pre-assembled sub-assemblies at the designated collaboration area. The final step involves a joint activity, where the cobot and participant mesh the sub-assemblies together. However, to ensure experimental control, the participants were asked to assemble one product at a time, i.e., completing the current production cycle before starting the next sub-assembly.

Layout

The layout of the experimental setup is same as Chapter 6, except three notable differences: pre-assembled sub-assemblies, fixed position of participant's component box, and dedicated experimenter space. The layout consisted of an L-shaped table divided into designated areas for the participant and the cobot, as depicted in Figure 7.3. The participant's side served as their dedicated assembly area, while the cobot's side held a readily accessible matrix of pre-assembled sub-assemblies. Like in the previous chapter, the cobot was positioned at the corner of the L-shaped table, allowing access to both the participant (for joint activity) and the sub-assemblies. A Logitech C920 Pro HD webcam was positioned approximately 1.5 meters in front of the participant.

In Chapter 6, the participants could choose where to keep the box with the assembly parts, potentially impacting production cycle time due to varying proximity. To mitigate



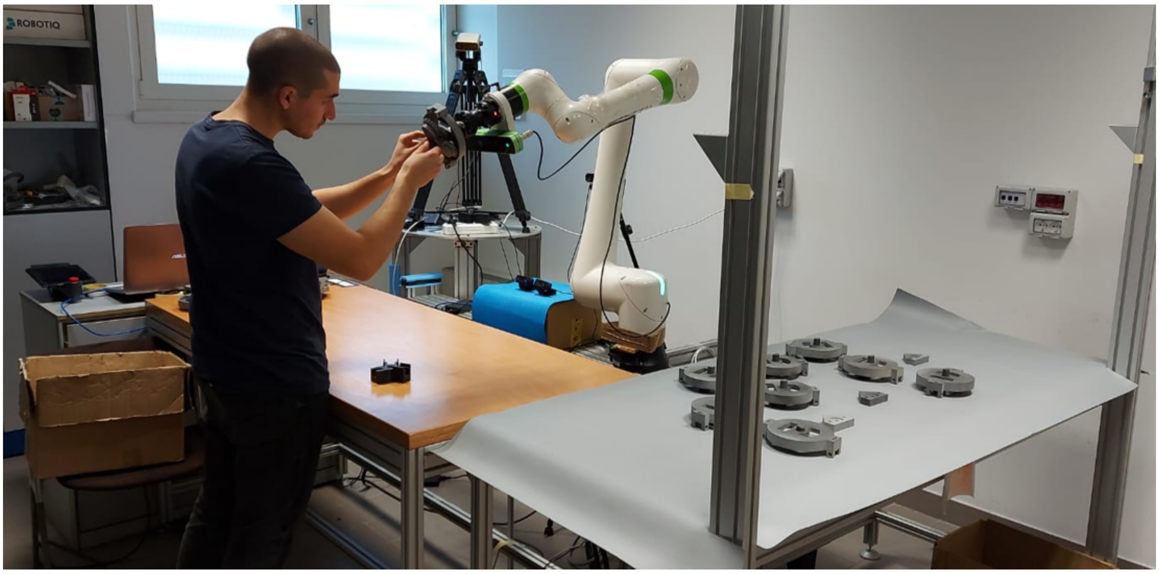


Figure 7.3: An overview of the experimental setup. On the top is an illustration of the layout of the setup, which is similar to Chapter 6 with some minor differences. On the bottom is the side view of a participant engaged in the assembly task. The dedicated front camera records the participant's face video during the experimental sessions. Additionally, a chest band worn by the participant (not visible) records ECG data. The layout illustration was created specially for this thesis; the copyright of participant's side view (bottom image) remains with the authors [Prajod et al., 2024a].

this factor in the current study, the box was positioned fixed at the left of each participant, ensuring consistency and minimizing timing variations arising from individual placement choices.

A dedicated area was designated for the experimenter on the left side of the participant's workspace (see Figure 7.3). This location provided a full view of the setup, allowing the experimenter to monitor the participant's progress and the cobot's activity. Additionally, the experimenter had access to a laptop for controlling the cobot's production rate in real-time.

Cobot capabilities

During the study, the participants collaborated with a Fanuc CRX-10iA/L cobot, an industrial robotic arm with a payload of 10 kg. The cobot was equipped with a Pickit3D camera for locating pre-assembled sub-assemblies within its workspace. It also utilized a Robotiq Hand-e gripper to pick up and manipulate the sub-assemblies during the collaborative assembly task. Within each production cycle, the cobot assists the participant by presenting a pre-assembled part in a convenient position for the final collaborative activity: meshing sub-assemblies. The cobot was programmed to release the sub-assembly only when the participant pressed a foot switch. Additionally, the cobot performed a scanning motion over the sub-assemblies when it was not actively engaged in the collaborative assembly.

7.3.2 Controlled Conditions

The insights gained from Chapter 6 revealed three distinct scenarios based on the production rates of the participant and the cobot: participant waiting for the cobot, cobot waiting for the participant, and synchronized assembly. These scenarios were translated into three distinct challenge levels of the assembly task, designed to evoke the states in the three-channel Flow model. Cobot behavior was modified to achieve these three challenge levels, as described below.

1. **Slow condition**: During this condition, the cobot performed a scanning motion using its wrist-mounted camera over the sub-assembly matrix before selecting and delivering one to the participant. This process extended the production cycle, resulting in an average time of 55 seconds from the start of the cycle to the cobot being ready for the joint activity.

2. **Fast condition**: To create a contrasting experience, the scanning motion was eliminated in the fast condition. The cobot proceeded directly to the next sub-assembly, picked it up, and delivered it to the participant. This approach significantly reduced the production cycle time, with an average of only 15 seconds before the cobot was ready for the joint activity.

3. **Adaptive condition**: This condition utilized a "Wizard of Oz" methodology. The cobot initially performed the scanning motion, but the timing of sub-assembly delivery was controlled by the experimenter acting as a wizard. The wizard triggered the cobot to deliver a sub-assembly only when they deemed the participant was nearly

ready for the joint activity. This eliminated a fixed timing for the cobot. This condition aimed to synchronize the cobot's delivery with the participant's progress, dynamically adapting the production rate based on the participant's pace.

7.4 Validating Imbalance Conditions — A Pilot Study

Due to the longer delivery time, the Slow condition is expected to create periods of waiting and potentially lead to boredom. Conversely, the Fast condition's reduced delivery time is anticipated to create a demanding scenario, potentially inducing anxiety from perceived time pressure and inability to match the cobot's pace. In contrast, the Adaptive condition is designed to create an optimal challenge level by dynamically adjusting the cobot's production rate based on the participant's pace.

To validate the experience of boredom and anxiety, a pilot study was conducted with four participants. The study took place at Consiglio Nazionale delle Ricerche (CNR) - Lecco, Italy, and participants were recruited from the institute's campus.

7.4.1 Scenario

The pilot study began by introducing participants to the assembly task and providing them with the necessary training. Each experimental session lasted approximately 50 minutes and was divided into two phases. During the first phase, participants collaborated with the cobot under the Slow condition. In the subsequent phase, the cobot functioned under the Fast condition. To induce a more genuine response and authenticity, the transition between these phases involved a simulated system failure. The cobot paused at the end of the first phase, prompting the experimenter to respond to this "fake failure" by adjusting the setup. Participants were then asked to adapt to the new pace of the cobot. The two phases were recorded using a 1080p camera.

In the pilot study, validation was limited to the Slow and Fast conditions. This decision stemmed from the empirical observation that states such as boredom and anxiety/stress are usually triggered more quickly and within shorter sessions. Conversely, while achieving a balance between challenge and skill is known to be conducive to flow, it does not ensure the elicitation of flow.

7.4.2 Questionnaires and Interview

After the two phases, the participant moved to a separate room to review six video clips from the experimental session. For each clip, participants completed two questionnaires: the Semantic Differential [Mehrabian and Russell, 1974] for emotion assessment and the Flow short scale [Rheinberg et al., 2003; Rheinberg, 2015]. The semantic differential scale measured emotional experience in terms of valence, arousal, and dominance. Each dimension was assessed using six pairs of adjectives, with nine spaces between each pair. The Flow short scale measures the flow experience using 10 components. Each component was rated on a 7-point Likert scale from "strongly disagree" to "strongly agree". A high total score suggests an intense experience of flow.

Category	Valence	Arousal	Dominance
Flow	+	+	+
Awe	+	+	−
Relaxed	+	−	+
Hopeful	+	−	−
Hostile	−	+	+
Anxious	−	+	−
U-Boredom/Disdain	−	−	+
O-Boredom/Apathy	−	−	−

Table 7.3: Categories used in coding interview data and their corresponding emotion dimensions (Valence, Arousal, Dominance)

Additionally, a semi-structured interview was conducted, where the participants shared the emotions experienced during the experimental session. The interview data was analyzed using the deductive category assignment method [Mayring, 2014]. Eight emotion categories were established by considering emotion dimensions (see Table 7.3), and these categories were employed for coding the interview data. The questionnaire and interview session lasted for approximately 45 minutes.

7.4.3 Analysis and Insights

Questionnaire Evaluation

The normalized valence, arousal, dominance, and flow values obtained from the questionnaires for each participant are presented in Figure 7.4. Some participants' flow experiences were more closely tied to specific emotional dimensions such as valence or arousal, while others showed correlations with multiple emotional dimensions. Specifically, for Participant 1, valence and arousal appear to track the flow trend. Meanwhile, for Participant 2, valence appears to best align with the flow trends. In contrast, for Participant 4, arousal seems to most closely follow the flow trends. Although the trend following is not robust for Participant 3, there appears to be a slight alignment with the flow trend, particularly for arousal and dominance.

Interview Evaluation

In the pilot study, Participant 1 experienced a state of relaxation during the Slow condition but encountered a range of emotions, including anxiety and hostility, during the Fast condition. They expressed feeling stressed due to the perceived responsibility for slowing down the assembly. Eventually, in the Fast condition, Participant 1 adjusted their pace, stating, "*It's enough for it to be repetitive; I don't need it also to be to be fast*".

Participant 2 described feelings of relaxation and u-boredom during the Slow phase but mentioned being distracted, leading to assembly mistakes. Transitioning to the Fast phase,

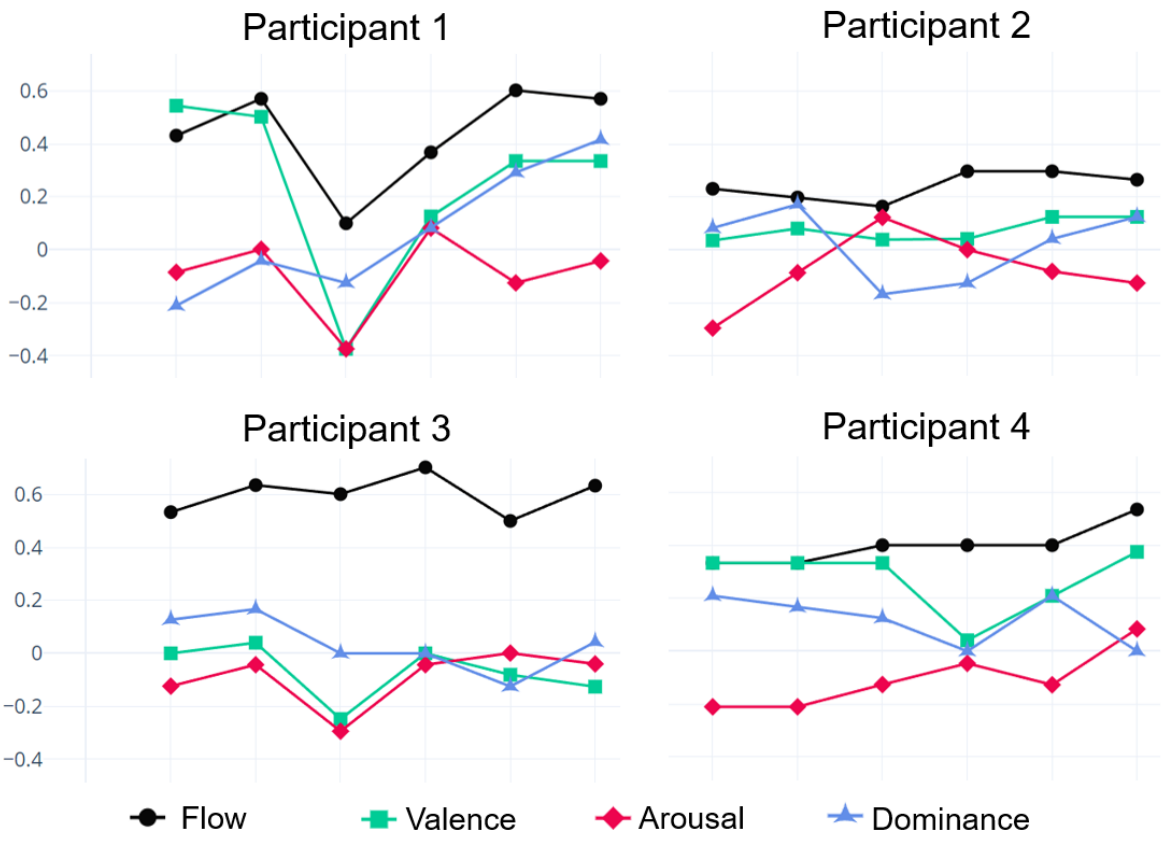


Figure 7.4: Plots illustrating the flow, valence, arousal, and dominance values acquired from the questionnaires for each participant. Each data point on the graph represents the respective dimension value for a working phase clip, resulting in a total of six points per participant for each dimension.

the participant felt stressed by the cobot's production rate. They reported thinking "*[...] this is my job and the robot is just a robot. So he has no feeling; he can wait. [...] also I had to do the most difficult part like matching the gears, with the clips and stuff*". This reappraisal of the situation led them to eventually operate at their own pace.

Similar to Participant 2, Participant 3 initially felt relaxed in the Slow condition, then got distracted, and later experienced boredom. In the Fast condition, they did not increase their production speed and reported that they "*didn't care anymore*" about the task. The participant could not identify a reason for this feeling of indifference.

Similar to other participants, Participant 4 also reported feeling relaxed in the Slow condition. They did not feel rushed in the Fast condition because they thought "*since the number of movements that the machine had to do was smaller than what I had to do, [cobot] was just completing the task earlier than me*". The participant placed blame on the task giver rather than themselves, expressing potential anger in real situations ("*[...] especially if it was in a real situation then I would have thought that the whole planning of the operation was bad because even if I strived, I wouldn't have managed to be so quick and be always on time for the robot. So yeah, I would be angry in case it would always be like that*").

Insights

During the interview, the participants commonly reported feelings of relaxation (4 out of 4) and boredom (2 out of 4) in the Slow condition. Conversely, stress (2 out of 4), anxiety (1 out of 4), hostility (1 out of 4), and anger (1 out of 4) were associated with the Fast condition. The interview data suggests that the Slow condition tends to evoke negative arousal states such as relaxation and boredom, while the Fast condition elicits high arousal states like stress and anxiety. Additionally, the Slow condition generally evokes positive dominance emotions, while the Fast condition evokes negative dominance emotions. However, while the Fast condition predominantly induces negative valence states, there is no clear pattern for the Slow condition.

Participants tended to adjust their task pace after re-evaluating the situation during the Fast condition, opting to operate at their own rhythm. This pattern is consistent with observations in the flow plots (refer to Figure 7.4), where a higher level of flow is noted during the later stages of the experimental session. This observation supports the inclusion of the Adaptive condition, as it suggests that tailoring the collaboration to the participant's production rate may enhance their experience of flow.

7.5 Dataset Collection

Following the pilot study, the experimental setup was leveraged to gather data for analyzing emotional and HRV responses across the three experimental conditions. Subsequently, this data was utilized to develop models aimed at facilitating flow experiences in industrial settings. A total of 37 adult volunteers participated in the study, consisting of 8 females and 29 males, with ages ranging from 18 to 48 years (mean = 29.03, SD = 7.08). Participants were predominantly students and staff from National Research Council of Italy - Lecco campus, with the majority being Italians (33 Italians, 4 non-Europeans). Recruitment was conducted through word-of-mouth and advertisements in public areas. Notably, none of the participants had prior experience working with an industrial cobot.

7.5.1 Study Protocol

The experiment began with a preparatory phase, during which participants were introduced to the setup and task requirements. They received detailed information about data collection procedures and were invited to provide informed consent, with the assurance that they could opt out of the experiment at any time. Following consent, participants provided demographic details and were instructed to wear a chest band for ECG data collection. During this phase, participants were encouraged to practice the task until they felt comfortable with the assembly steps. An experimenter was present in the room throughout the experiment to address any technical issues and acted as the “Wizard" in the Adaptive condition. Fluent in both Italian and English, the experimenter provided instructions and questionnaires in the participant's preferred language. Interaction between the experimenter and participants was kept to a minimum, and participants were instructed to refrain from providing feedback until the conclusion of the experiment.

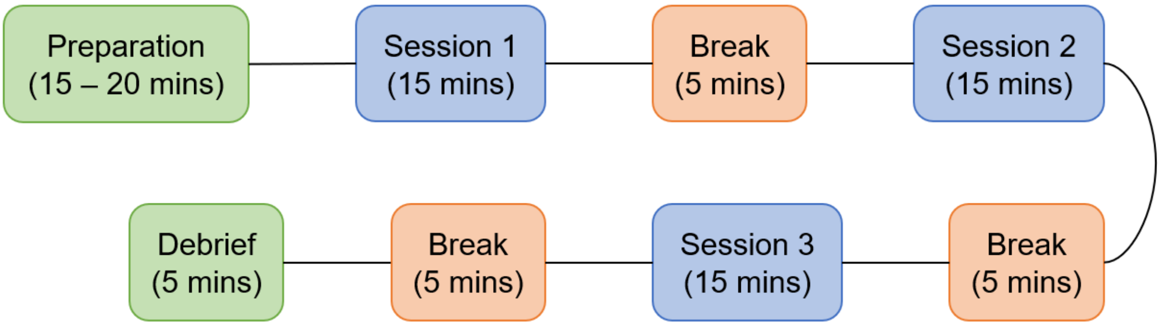


Figure 7.5: An overview of the experimental protocol consisting of three conditions (Slow, Fast, Adaptive). To mitigate any ordering effect, the condition sequences were counterbalanced among participants. The copyright remains with the authors [Prajod et al., 2024a].

As illustrated in Figure 7.5, the study employed a within-subjects design, where each participant engaged in all three experimental conditions, separated by 5-minute breaks between consecutive sessions. Each session had a duration of 15 minutes, during which participants were involved in the continuous assembly of gearboxes alongside the cobot. The sequence in which the three conditions were presented was randomized and balanced to minimize any potential effects related to the order. During the breaks, participants filled out questionnaires regarding their experience in the preceding session. After completing all three sessions, participants were debriefed about the experiment. The study was conducted according to the guidelines of the Declaration of Helsinki and approved by Commissione per l'Etica e l'Integrità nella Ricerca of the National Research Council of Italy (protocol n. 0085720/2022 of 23/11/2022).

7.5.2 Data Acquisition Tools

After each condition, three questionnaires - NASA-TLX (Task Load IndeX), SAM (Self-Assessment Manikin), ELoC (Experiential Locus of Control) - were administered to investigate how the task load and experiences differed across the experimental conditions. Additionally, videos and ECG data were collected during the experimental sessions.

NASA-TLX

The NASA-TLX questionnaire [Hart and Staveland, 1988] includes six sub-scales that capture workload factors: Mental Demand, Physical Demand, Temporal Demand, Frustration, Effort, and Performance. Participants rate each sub-scale on a scale ranging from 1 (very low) to 20 (very high). This questionnaire was selected due to its ability to assess the type of task load experienced by participants. As highlighted in Section 7.2.3, prior studies in the literature have primarily focused on tasks with high mental demands. However, in this case, the increased production rate of the cobot is expected to result in a higher number of completed assemblies. Therefore, it is anticipated that the task load will primarily be attributed to physical and temporal demands.

SAM

The SAM questionnaire [Bradley and Lang, 1994] is widely employed to assess emotional responses and subjective experiences. It utilizes pictorial representations to assess three dimensions of emotion: Valence, Arousal, and Dominance. Participants provide ratings on 9-point Likert scales reflecting these dimensions. The Valence scale ranges from positive to negative, while the Arousal scale spans from excitement to calmness, and the Dominance scale ranges from low to high control. This questionnaire is a non-linguistic alternative to the Semantic differential tool used in the pilot study [Bradley and Lang, 1994].

ELoC

The ELoC questionnaire [Jang et al., 2016] is designed to assess an individual's perception of control over their experiences and circumstances. It consists of three sub-scales adapted from the Internal Control Index tool [Duttweiler, 1984], which measures the locus of control of an individual. Each sub-scale is rated on a 5-point Likert scale, where 1 corresponds to “Completely Disagree" and 5 to “Completely Agree". The total score obtained on the ELoC questionnaire can range from 3 to 15, where a higher score indicates a stronger internal control. This questionnaire was included in the data collection process because the sense of control is considered an important indicator of flow.

Sensors - Videos and ECG

The videos of participants were captured using a frontal camera, recording at a resolution of 1920 × 1080 pixels and a frame rate of 25 fps. An SSI [Wagner et al., 2013] pipeline was employed to capture and store the videos.

The participants were instructed to wear a Polar H10 chest band to record their ECG signals at a rate of 130 Hz. This chest band was connected wirelessly to an Android phone via Bluetooth, enabling the signals to be received and stored using an SSJ pipeline [Damian et al., 2018].

7.6 Analysis of Questionnaires

The participants' subjective workload and experience were measured using NASA-TLX, SAM, and ELoC questionnaires. Table 7.4 summarizes the mean response values for the components of the questionnaires across the three conditions.

7.6.1 NASA-TLX

The subjective workload factors experienced by participants were assessed using the NASA-TLX questionnaire. Analysis (see Table 7.4) revealed that the Fast condition, resulted in higher workload scores for Effort, Mental demand, Physical demand, and Temporal demand. Conversely, the Slow condition yielded the lowest perceived workload in these categories. The differences in perceived temporal demand were the most pronounced

Category	Slow	Fast	Adaptive
NASA-TLX (max 20)			
Mental demand	4.81	**6.35**	4.95
Physical demand	4.59	**6.84**	5.05
Temporal demand	4.73	**10.54**	6.08
Effort	5.16	**7.76**	5.65
Performance	**7.30**	6.81	6.81
Frustration	**5.35**	**5.35**	4.27
SAM (max 9)			
Valence	**4.29**	4.03	3.84
Arousal	**6.84**	6.35	6.77
Dominance	**7.19**	7.00	7.03
ELoC (max 15)			
ELoC total	9.48	**9.58**	9.52

Table 7.4: The average responses to the NASA-TLX (on a 20-point scale), SAM (on a 9-point scale), and ELoC (total 3 - 15) questionnaires recorded after each conditions

across the three experimental conditions. This observation aligns with the initial expectations regarding temporal and physical demands being impacted by the cobot's production rate. However, it is noteworthy that the cobot's influence extended to other workload dimensions as well.

Despite completing the most assemblies in the Fast condition, participants reported the highest perceived performance in the Slow condition. This suggests a potential trade-off between speed and accuracy or perceived competence. Additionally, the Adaptive condition stood out with the lowest frustration score, indicating a potential benefit of tailoring the cobot's production rhythm to individual needs.

7.6.2 SAM and EloC

The SAM questionnaire assessed participants' emotional valence, arousal, and dominance across the three conditions (see Table 7.4). Notably, dominance scores remained high across all conditions, suggesting participants generally felt in control of their emotional experience. This implies the sessions did not induce significant instances of low-dominance emotions like anxiety or apathy. Furthermore, consistently elevated arousal levels were observed across conditions. While the specific reason for this remains unclear, it potentially reflects a certain level of task engagement among participants. Despite slightly lower average valence scores, these fall close to the absolute scale midpoint, indicating a more neutral state rather than any significant negative emotions.

Interestingly, the questionnaire responses revealed no considerable differences in emotional ratings between conditions. This suggests that despite varying workloads induced by the cobot's production rate, participants' overall emotional experiences remained comparable.

Similarly, the ELoC scores exhibited no notable variations across conditions (see Table 7.4). This indicates that the participants' perceived control over their actions and the task remained relatively stable regardless of the cobot's rhythm.

While varying cobot assistance impacted workload perception, these findings suggest that such influences may not have directly translated into strong changes in participants' overall emotional state or perceived control during the task.

7.7 Analysis of Emotion Estimation

The recorded facial videos were analyzed to assess potential variations in emotional responses across the three experimental conditions. This analysis employed an SSI pipeline designed to process individual video frames and infer emotional states. First, face regions were identified within each frame using MediaPipe's BlazeFace detection model [Bazarevsky et al., 2019]. Then, these detected regions were cropped and fed as input to a deep-learning model trained for the emotion estimation task.

7.7.1 Emotion Estimation Model - Valence and Arousal

The analysis employed a convolutional neural network to estimate emotions based on participants' facial expressions. This model categorized images into seven discrete emotion classes (Neutral, Happy, Sad, Surprise, Fear, Disgust, and Anger) and additionally provided continuous valence and arousal scores ranging from -1 to 1.

Dataset

The model was trained on the AffectNet dataset, described in detail in Section 4.3.1. To ensure the reliability of data, the dataset underwent cleaning based on the pre-processing techniques presented by Toisoul et al. [2021]. This process involved removing images where the assigned emotion class did not align with the corresponding valence-arousal values. This refined dataset contained around 220K images categorized into seven emotion classes, along with valid valence-arousal values. The data was further split, with 85% allocated for training the model and 15% reserved for evaluation.

Training

The model architecture leveraged a pre-trained VGG16 network followed by a fully connected layer and three output layers. Each output layer utilized specific activation functions: Softmax for classifying emotions into discrete categories, and Tanh for estimating continuous valence and arousal scores. All images were resized to 224 × 224 pixels and augmented through techniques such as width shift, height shift, zoom, and horizontal flip.

The model architecture and training approach were adapted from Section 4.3.1, due to their demonstrated ability to learn facial action units, which are crucial for accurate emotion recognition. The model was trained using Stochastic Gradient Descent (SGD) optimization with an initial learning rate of 0.001. This learning rate was dynamically adjusted by reducing it by 10% after every 70,000 training steps. The model training utilized focal loss [Lin et al., 2017] for emotion classification and shake-shake loss [Toisoul et al., 2021] for valence and arousal predictions. To prevent overfitting during training, early stopping with patience of five epochs was implemented. This technique halted training when the validation loss stopped improving.

Evaluation

On the AffectNet test set, the model achieved an accuracy and F1-score of 76% for the discrete emotion classification task. This performance aligns with previous studies utilizing the AffectNet dataset, as reported by Mollahosseini et al. [2017] and Toisoul et al. [2021].

Concordance Correlation Coefficient (CCC), Root Mean Squared Error (RMSE), and Sign Agreement (SAGR) were employed to assess the model's ability to predict continuous emotional dimensions, consistent with prior research. For valence prediction, the model achieved a CCC of 0.852, RMSE of 0.266, and SAGR of 83.1%. Similarly, for arousal prediction, the model achieved a CCC of 0.763, RMSE of 0.277, and SAGR of 81.2%. The model's performance on both valence and arousal prediction was comparable to state-of-the-art methods.

Post-processing

The pre-trained emotion recognition model was applied to the recorded videos from each experimental condition. This process generated a sequence of emotion estimations for each frame, providing a nuanced characterization of emotional responses throughout each session. To ensure the reliability of these estimations, frames where no face was detected were excluded from further analysis. Additionally, as participants may require some time to adjust to the robot's behavior and fully engage in the task, emotion estimations from the initial five minutes of each session were excluded.

7.7.2 Are Valence and Arousal Indicators of Flow?

This analysis focused primarily on valence and arousal values, as they offer a more nuanced representation of dynamic emotional states compared to discrete categories. For each experimental condition, the mean valence and arousal were computed across all frames to capture overall emotional trends. This session-wise mean valence and arousal were computed for each participant.

Overall patterns

Averaged across participants, mean valence levels were lowest for the Slow condition (-0.025) and highest for the Fast condition (-0.018). Similar trends were observed in mean arousal levels, with Slow (0.053) being the lowest and Fast (0.074) being the highest. The Adaptive condition (valence: -0.023, arousal: 0.071) exhibited intermediate values in both dimensions.

Statistical analysis

A repeated measures ANOVA test was conducted to determine statistically significant differences in mean valence and arousal across conditions. Both measures satisfied the assumptions of normality and homogeneity of variance. The analysis revealed a significant difference in mean arousal ($F = 8.23$, $p < 0.001$) but not in mean valence.

Further pairwise comparisons with Holm correction identified significant differences in mean arousal between the Slow condition and both Fast ($p = 0.012$) and Adaptive ($p = 0.015$) conditions. No significant difference was found between Fast and Adaptive conditions ($p = 0.884$).

Interpretation

While significant differences were detected in arousal, it's worth noting that the observed mean values across all conditions fell within the neutral range (0 to 0.1). This observation suggests that participant facial expressions might not necessarily be a reliable indicator of perceived challenge levels in this case.

7.7.3 Insights and Discussion

The study anticipated negative emotions in participants for the Slow and Fast conditions due to the imbalance in challenge and skill. Conversely, the Adaptive condition was expected to facilitate positive emotions associated with the flow state. However, analysis revealed no significant difference in valence levels across the three conditions. Interestingly, average valence and arousal across all conditions remained close to zero, indicating a predominantly neutral emotional state throughout the experiment. This finding suggests that the robot's programmed behavior did not elicit strong negative emotions from participants.

The results of the SAM questionnaire also revealed no significant differences in self-reported valence, arousal, or dominance across the three challenge conditions (see Section 7.6.2). This indicates that participants' emotional experiences remained relatively consistent throughout the experiment, regardless of the task difficulty level.

Burns and Tulip [2017] explored the use of facial emotion estimation (valence and arousal) in a gaming context for dynamically adjusting the difficulty level of the game. They observed that players' facial expressions remained near neutral for extended periods, with occasional brief spikes. This aligns with the findings of the current study, suggesting limited emotional expression by participants. In line with Burns and Tulip, this study also suggests that facial expressions alone might not be sufficient for dynamically adapting the cobot's behavior.

7.8 Analysis of HRV

This analysis leveraged the HRV features derived from participants' ECG data collected during the experimental sessions. The same set of 22 HRV features (time domain, frequency domain, and Poincaré plots) detailed in Chapter 5 were extracted. All features extracted from these domains were included in the analysis.

7.8.1 HRV Feature Extraction

To ensure accurate and reliable analysis, the ECG signals underwent the following series of cleaning steps prior to feature extraction.

- **Noise removal**: A second-order Butterworth band-pass filter with a cut-off frequency of 8-20 Hz was applied to remove noise in the signal, following recommendations from previous research [Elgendi et al., 2010].
- **Heartbeat detection**: The ECG signals were segmented into 1-minute intervals using a sliding window with 1-second shifts. Heartbeats were identified within each segment using the method proposed by Elgendi et al. [2010]
- **Segment exclusion**: Segments with missing beats, false detections (excessive beats), or heart rates outside the range of 50-180 beats per minute were excluded. Segments were deemed invalid if the time between consecutive beats exceeded 1200 milliseconds (indicating a heart rate less than 50 bpm). Similarly, segments with consecutive beats occurring less than 333.33 milliseconds apart (indicating a heart rate greater than 180 bpm) were also excluded.

Similar to emotion analysis from the previous section, data from the initial 5 minutes of each session was excluded. Additionally, participants with less than 5 minutes of clean ECG data in each session were excluded from further analysis. This criterion led to the exclusion of nine participants.

7.8.2 Are HRV Features Indicators of Flow?

This section details the analysis procedure for HRV features using heart rate as a representative example. Measured in beats per minute, heart rate is one of the commonly used HRV features due to its established relationship with arousal levels [Rissler et al., 2020]. This analysis procedure was applied to each of the 22 HRV features.

Overall patterns

To account for inter-individual variability in physiological responses, the MinMax normalization was applied to heart rate and HRV features for each participant. This ensured that all features were scaled to a common range (0 - 1), mitigating potential biases due to individual baseline differences.

The normalized heart rate data for the three experimental conditions (Slow, Fast, Adaptive) are presented in Figure 7.6 (top) as box plots. The Fast condition exhibits the highest average normalized heart rate (mean = 0.554), followed by the Adaptive condition (mean = 0.485), and lastly, the Slow condition (mean = 0.402).

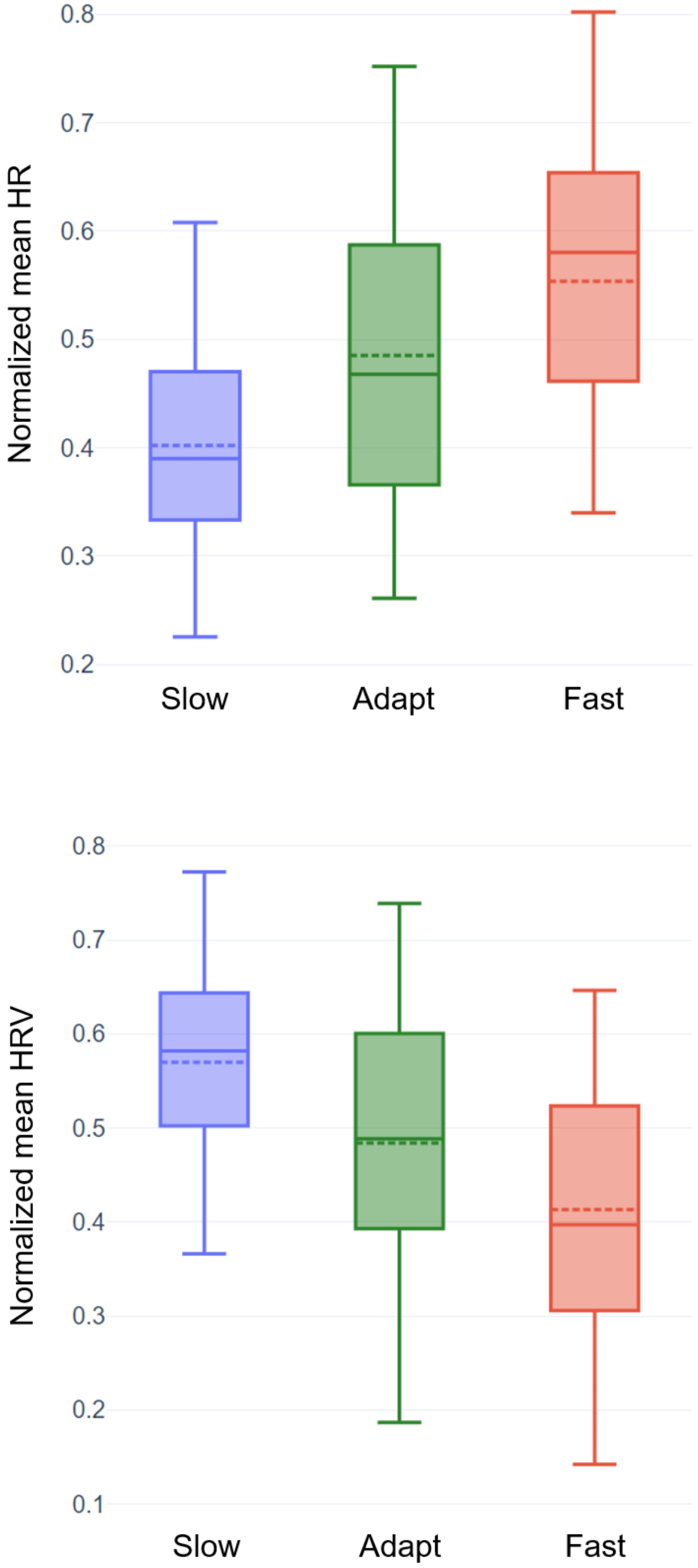


Figure 7.6: Box plots of the normalized mean heart rates (top) and normalized mean HRV (bottom) in the three conditions. Each box plot summarizes the distribution within a condition. The dotted line represents the mean of the distribution.

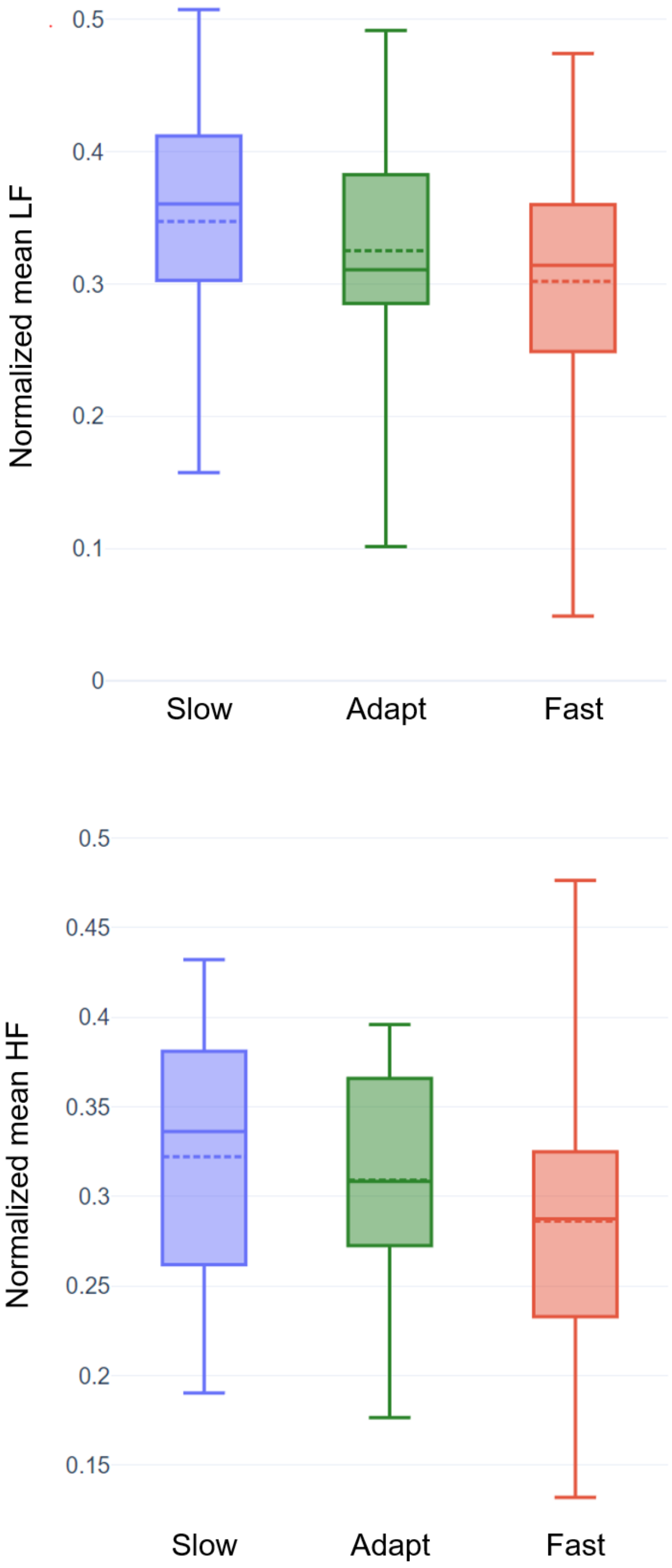


Figure 7.7: Box plots of the normalized mean LF (top) and normalized mean HF (bottom) in the three conditions. Each box plot summarizes the distribution within a condition. The dotted line represents the mean of the distribution.

Other commonly studied HRV features, including mean HRV, LF, and HF components, were also analyzed and are presented in Figures 7.6 (bottom), 7.7 (top), and 7.7 (bottom), respectively. The trends in these figures align with the established literature, demonstrating an increase in heart rate and a decrease in HRV with increasing challenge levels. As expected, the Adaptive condition, characterized by balanced challenge and skill, exhibited a relatively moderate heart rate and HRV. Both LF and HF components followed a decreasing trend with increasing challenge, with the Slow condition showing the highest values and the Fast condition showing the lowest.

Statistical analysis

Feature	ANOVA	Post hoc	C1 vs. C2	C1 vs. C3	C2 vs. C3
Time domain features					
HR	< 0.001 *	= 0.002 *	< 0.001 *	= 0.038 *	= 0.056
Mean NN	< 0.001 *	= 0.002 *	< 0.001 *	= 0.038 *	= 0.054
SD NN	= 0.017 *	= 0.231			
CV NN	= 0.553	= 0.553			
Med NN	< 0.001 *	< 0.001 *	< 0.001 *	= 0.018 *	= 0.027 *
Mad NN	< 0.001 *	= 0.018 *	< 0.001 *	= 0.046 *	= 0.052
RMSSD	= 0.123	= 0.862			
SDSD	= 0.121	> 0.99			
IQR NN	= 0.005 *	= 0.082			
pNN50	= 0.026 *	= 0.255			
pNN20	< 0.001 *	= 0.005 *	< 0.001 *	= 0.146	= 0.012 *
TI NN	= 0.552	> 0.99			
TI	< 0.001 *	= 0.002 *	< 0.001 *	= 0.002 *	= 0.595

Table 7.5: Significance test results (p-values) for the time domain HRV features. The * *symbol next to the p-values indicates that the result is considered statistically significant (< 0.05).*

Similar to the analysis approach adopted for emotion estimation, a repeated-measures ANOVA was conducted on the average normalized heart rate values across the three experimental conditions. Before the analysis, checks were conducted to ensure the data met the assumptions of homogeneity of variance and normality.

The ANOVA test revealed a statistically significant difference between at least two of the experimental conditions (F = 10.59, p < 0.01), indicating a considerable impact of condition on normalized heart rate values. Subsequent post hoc pairwise t-tests with Holm correction identified a significant difference between the Slow condition and both the Adaptive (p = 0.038) and Fast conditions (p < 0.001). Notably, the difference in average heart rate between the Fast and Adaptive conditions only obtained a p-value of 0.056, suggesting a potential trend.

Feature	ANOVA	Post hoc	C1 vs. C2	C1 vs. C3	C2 vs. C3
Frequency domain features					
LF	= 0.017 *	= 0.216			
HF	= 0.020 *	= 0.221			
LF/HF	= 0.334	> 0.99			
LF/Total	= 0.411	> 0.99			
HF/Total	= 0.262	> 0.99			
Poincaré plot features					
SD1	= 0.121	= 0.968			
SD2	= 0.012 *	= 0.183			
SD1/SD2	= 0.285	> 0.99			
S (Area)	= 0.018 *	= 0.217			

Table 7.6: Significance test results (p-values) for the frequency domain and non-linear (Poincaré plots) HRV features. The * symbol next to the p-values indicates that the result is considered statistically significant (< 0.05).

Tables 7.5 and 7.6 systematically present the outcomes of the statistical analysis applied to each HRV feature. The pair-wise testing was conducted only if the results were significant after the Holm correction.

Interpretation

The analysis of heart rate and HRV aligns with observations from previous studies (see Section 7.2.2) that primarily investigated mentally demanding tasks. This observation signifies that heart rate is a good indicator of flow even in other types of workloads.

Notably, several HRV features, especially those derived from the temporal domain, exhibited statistically significant differences across experimental conditions even after applying the Holm correction for multiple comparisons within the ANOVA test. These findings suggest that HRV features are promising indicators of perceived challenge levels within the context of human-robot collaboration tasks.

7.8.3 Insights and Discussion

The statistical results obtained for HRV in this study are consistent with those reported by Keller et al. [2011]. In their study, Keller et al. found a significant difference in mean HRV between the low-challenge (boredom) and higher-challenge (fit, anxiety) conditions. Similar to the findings here, their study found a trend-level significant ($p < 0.1$) difference between the Fit and Anxiety conditions.

Furthermore, the plots presented for LF and HF components resonate with the findings of Tozman et al. [2015] regarding participants' physiological responses during the three driving conditions (boring, fit, anxiety). The observed decrease in HF, typically associated with parasympathetic activity (relaxation), was anticipated.

Regarding LF, interpretations vary. Some studies associate it solely with sympathetic activation (arousal), while others view it as a combined measure of both branches [Malik et al., 1996]. Adopting the latter interpretation, the decreasing LF trend suggests a reduction in relaxation with increasing challenge. This aligns with the Adaptive condition exhibiting moderate levels of relaxation and arousal.

While Tozman et al. distinguished between three challenge levels using LF and HF components of HRV, the current analysis did not find statistical differences between all conditions. This discrepancy might be attributed to the differing methods used to induce challenge. Tozman et al. incorporated social evaluation as a stressor in the anxiety condition, likely triggering intensified physiological responses. In contrast, this study solely manipulated challenge through the robot's behavior, avoiding the introduction of external stressors.

7.9 Models For Challenge Adaptation

Recognizing participants' perceived challenge levels during human-robot collaboration is crucial for optimizing interaction and promoting flow experiences. While traditional methods often rely on self-reported data, physiological measures like HRV provide a continuous and non-obtrusive assessment of the operator's response to task demands [Peifer, 2012; Irshad et al., 2023]. This section explores the potential of HRV features in predicting challenge levels.

7.9.1 Challenge Prediction

The previous section revealed promising relationships between HRV features and the experimental conditions designed to manipulate the perceived challenge level. Building upon these findings, this chapter implemented machine learning models for predicting challenge levels from HRV data. This section presents two challenge prediction scenarios:

- **Multi-class prediction**: This model was trained to classify HRV data into three classes (Slow, Adaptive, Fast).
- **Binary-class prediction**: This model was developed to distinguish the Slow condition from the other two conditions. This approach was motivated by the non-significant differences observed in previous analyses (see Table 7.5) suggesting limited discernible variations in HRV responses between the Fast and Adaptive conditions. The data from both Fast and Adaptive conditions were merged for training the binary-class prediction model.

7.9.2 Training and Evaluation

The challenge prediction models followed a simple feed-forward neural network architecture. The network started with an input layer that received the pre-processed HRV

Model	Accuracy	F1-score
Baseline Multi-class (predicting one class)	0.333	0.167
Multi-class (low vs. fit vs. high challenge)	**0.493**	**0.459**
Baseline Binary-class (predicting majority class)	0.667	0.533
Binary-class (low vs. fit + high challenge)	**0.707**	**0.661**

Table 7.7: The average accuracy and F1 score achieved by binary-class and multi-class challenge prediction models during LOSO evaluation. The performance of baseline classifiers that always predict one class is also provided for comparison.

features. This was followed by two hidden layers, each containing 12 and 6 nodes respectively. To prevent overfitting and improve generalization, a dropout layer with a 10% rate was added after the input layer. This technique randomly dropped out a portion of neurons during training. The final output layer, depending on the prediction scenario (multi-class or binary-class), contained either 2 or 3 nodes. The Softmax activation function is applied in this layer.

Training of the models occurred in mini-batches of 128 samples using an SGD optimizer with a learning rate of 0.01. To ensure the models performed well on unseen data, a LOSO method was adopted for training and evaluating the models. The performances of these challenge prediction models were assessed in terms of accuracy and F1-score, and are presented in Table 7.7.

7.9.3 Insights and Discussion

The binary-class challenge prediction model outperformed the multi-class model. This outcome aligns with the HRV analysis in the previous section, suggesting potential challenges in distinguishing Fast and Adaptive conditions due to the similarity in their elicited physiological responses. Although the accuracy of the binary-class model might seem moderate for a binary classifier, it is comparable to existing binary flow detection models from the literature (see Section 7.2.3). In other words, the performance is similar to scenarios with mentally demanding tasks, indicating that moderate accuracy is not due to the physical or temporal demands of the HRC task. This implies that using HRV features to distinguish between low and high challenge levels in HRC settings is a viable approach.

Interestingly, the pilot study revealed that participants tended to adapt to faster cobots by adjusting their pace, potentially explaining the lack of significant differences between Fast and Adaptive conditions. Conversely, slower cobots led to distraction and assembly mistakes. It is critical to avoid assembly mistakes as they can significantly impact the production output [Klein et al., 2024]. This highlights the importance of detecting low-challenge situations in industrial settings to adapt the cobot's behavior accordingly. The binary-class model presents a promising step towards this goal.

7.10 Reflections and Remarks

This chapter addresses the topic of human-centered dynamic workload adaptations in industrial HRC. Specifically, it attempts to answer the question: Is it possible to detect if an operator feels under-challenged or over-challenged during an HRC task? Ideally, the task challenge is balanced to the skill level of the operator, a working condition conducive to the experience of flow.

While previous research has studied how the flow state manifests in mentally demanding tasks, there is a lack of studies investigating the flow experience in the HRC context. This investigation is necessary because an industrial HRC task often involves physical or temporal demands, rather than pure mental load. This chapter manipulated physical and temporal demands by varying a cobot's production rate, simulating under-challenged, over-challenged, and balanced "fit" conditions in an industrial HRC setting. During pilot testing, a crucial disadvantage of challenge-skill imbalance became evident: low-challenge conditions often led to distractions and eventually, assembly errors.

This chapter explored the potential of facial expressions and HRV features as indicators of perceived challenge level. Notably, HRV features were found to be reliable indicators of perceived challenge level. Subsequently, a model was trained to automatically detect the challenge level, which holds the potential to dynamically adapt the cobot's behavior.

It's worth noting that HRV features were previously employed for stress detection in Chapter 5. This chapter builds upon that work by demonstrating the extended utility of HRV features for detecting the perceived challenge level. Given that both stress and flow states can occur in industrial scenarios, the development of a combined model becomes a desirable future direction.

Chapter 8

Gaze to Initiate Collaboration in Industrial HRC

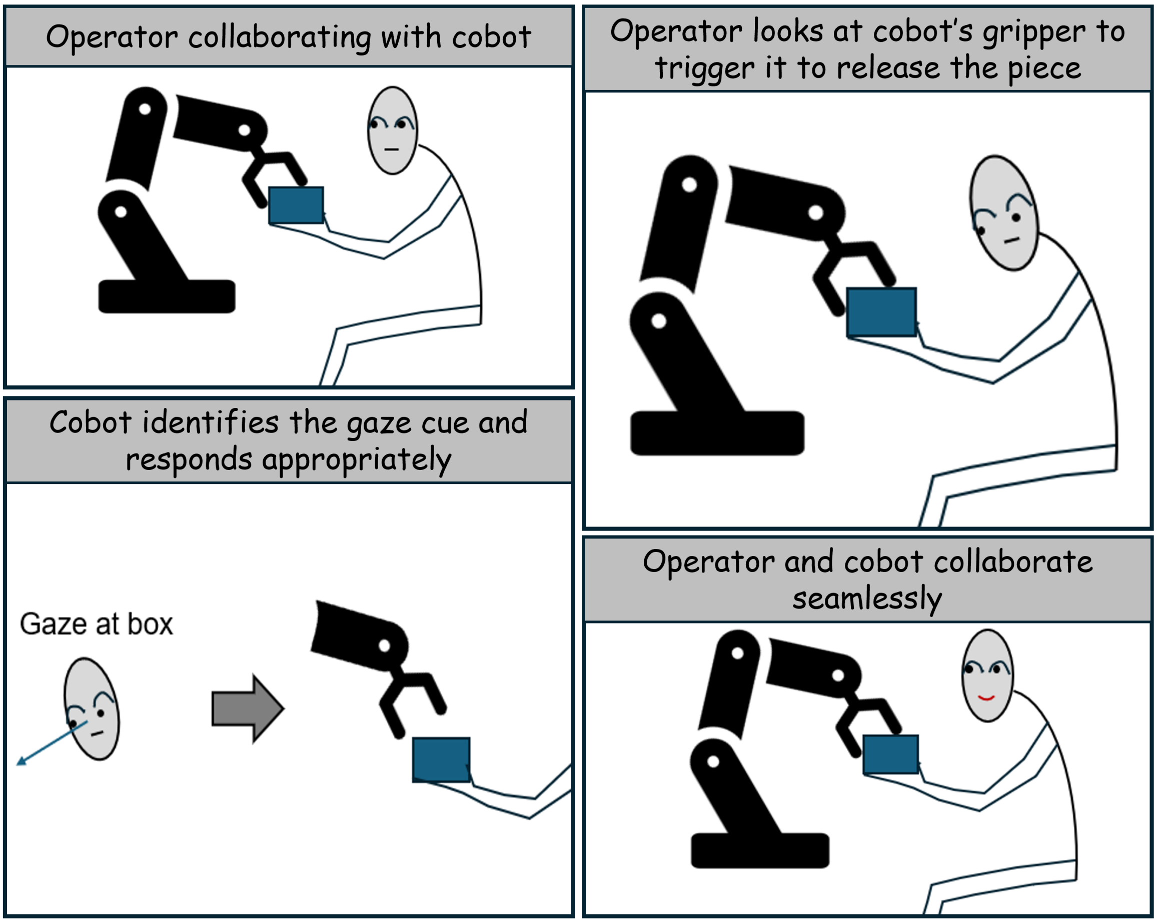

Operator collaborating with cobot
Operator looks at cobot's gripper to trigger it to release the piece
Cobot identifies the gaze cue and responds appropriately
Gaze at box
Operator and cobot collaborate seamlessly

Figure 8.1: A comic strip illustration of an operator communicating with the cobot using natural and intuitive gaze cues in an industrial scenario. During the task, the cobot recognizes the gaze cues of the operator and responds by performing the expected actions. This leads to a seamless collaboration between the cobot and the operator.

8.1 Overview

In industrial settings, cobots are transitioning from tools to collaborative partners. To achieve seamless collaboration, robots must not only be efficient, but also understand natural human cues, including gestures, voice commands, and even gaze patterns [Bauer et al., 2008; Glasauer et al., 2010; Buss et al., 2011; Gleeson et al., 2013; Mutlu et al., 2016; Romat et al., 2016; Campeau-Lecours et al., 2018]. An example industrial scenario is illustrated as a comic strip in Figure 8.1. Cobots equipped to understand natural communication can improve user acceptance and trust [Villani et al., 2018; Strazdas et al., 2020; Andronas et al., 2021; Kalatzis et al., 2023]. When a cobot can interpret human communicative cues like gestures and gaze, it can respond in ways that are more natural and expected. Such responses make the interaction more comfortable and lead to cobots being perceived as trustworthy partners. Furthermore, intuitive interactions reduce the cognitive load on human operators, allowing them to focus better and feel less mentally strained [Villani et al., 2018; Hasnain et al., 2013; Kalatzis et al., 2023]. This, in turn, can lead to increased efficiency, reduced errors, and ultimately, a more positive and productive work environment.

A key aspect of industrial Human-Robot Collaboration (HRC) is a physical joint activity involving simultaneous manipulation of an object by a cobot and a human operator. During human-human collaborations, specific gaze patterns emerge, serving as crucial communication cues that synchronize their actions and ensure seamless collaboration [Admoni and Scassellati, 2017]. While these human gaze patterns can also be observed in human-robot collaborations involving humanoid robots [Mehlmann et al., 2014; Kurylo and Wilson, 2019; Palinko et al., 2016], it remains unclear whether these gaze patterns emerge when collaborating with a robot without human-like features. Consequently, the potential of utilizing gaze-based cues for a natural, intuitive, and seamless collaboration between human operators and cobots remains largely unexplored. This chapter investigates these research questions using two experiments. The contents of this chapter expand upon the research previously published in:

* P. Prajod, M. L. Nicora, M. Mondellini, G. Tauro, R. Vertechy, M. Malosio, and E. André. Gaze detection and analysis for initiating joint activity in industrial human-robot collaboration. *arXiv preprint arXiv:2312.06643*, 2023b

  [*I contributed significantly to the study design and formulation of the hypotheses. I also performed data processing and developed the machine learning models. Furthermore, I conducted the analysis and derived insights.*]

* M. Lavit Nicora, P. Prajod, M. Mondellini, G. Tauro, R. Vertechy, E. André, and M. Malosio. Gaze detection as a social cue to initiate natural human-robot collaboration in an assembly task. *Frontiers in Robotics and AI*, 11:1394379, 2024

  [*This is a follow-up paper to the above paper. I trained machine learning models and developed the real-time detection pipeline.*]

8.2 Background Literature and Previous Works

Although verbal communication is the most direct way of conveying information, the potential of employing speech in industrial settings is often limited because of the noisy environment. Hence, it is essential to explore other modalities and social cues to enhance the industrial HRC experience through natural and intuitive communication. In addition to speech, humans employ many natural non-verbal cues including facial expressions, gestures, etc., to convey information. Gaze is one such non-verbal channel of communication that conveys valuable information not only in human-human interactions but also in human-robot interactions. This section presents background literature on the role of gaze in human-human interactions, followed by a brief discussion on the previous works investigating human gaze in human-robot collaborations.

8.2.1 Gaze in Human-Human Interactions

From infancy, humans utilize gaze as a fundamental communication signal. Even seemingly simple acts like looking at someone or an object involve coordinated movements of the eyes, head, and body [Rosander, 2020]. As cognitive development progresses, gaze evolves into a powerful tool for intentional communication [Camaioni, 1992], playing a critical role in establishing social cues [Hamilton, 2016].

Investigations into the neural correlates of gaze reveal its intricate connection to social behavior. Senju and Johnson [2009] propose that perceived eye contact activates brain regions associated with social interaction. This emphasizes the relationship between eye contact and social actions, not only at the behavioral level but also from a neurobiological perspective.

Several researchers have highlighted the significance of gaze in initiating interaction. Research by Cary [1978] analyzing video recordings of strangers in a waiting room revealed that eye contact consistently preceded conversations. Furthermore, Senju and Csibra [2008] investigated the relationship between eye contact and gaze following in infants, specifically demonstrating that eye contact can act as a strong cue for infants to pay attention and initiate gaze following behavior. Similar gaze-initiation behaviors are mimicked in robots so that they can establish and respond to eye contact during scenarios such as conversations, collaboration, and narration [Admoni and Scassellati, 2017].

Ferri et al. [2011] conducted experiments where a participant offered food to another person sitting across the table. Their findings indicated that the information about the receiver's gaze influenced the effectiveness of the feeding gesture, suggesting that gaze plays a crucial role in conveying social affordance. Similarly, Innocenti et al. [2012] investigated the impact of gaze on a more subtle requesting gesture – experimenter grabbing an empty glass while the participant lifted a juice bottle to pour the juice. Even in the absence of any verbal communication, the study demonstrated that the receiver's gaze behavior influenced the effectiveness of the gesture. These studies point to the potential of gaze in communicating requests, which has been exploited in HRC designs(e.g., work by Palinko et al. [2016]).

The role of gaze has been widely investigated in the context of human-human conversations [Kendrick and Holler, 2017; Degutyte and Astell, 2021]. The seminal work by Kendon [1967] identified two functions of gaze direction in dyadic interactions: monitoring and turn regulation. His analysis suggests that the listener tends to look at the speaker for long with few instances of gaze aversion, whereas the speaker displayed both gaze at listener and gaze aversion equally. Specifically, the speaker averted their gaze in the beginning of the round, likely for focusing and planning their speech, and looked at the listener for monitoring the attention and yielding their turn. However, later studies like Rossano et al. [2009] suggest that Kendon's findings may not extend to interactions that require multiple sequences to complete (e.g., question-answer). In these scenarios, gazing at the partner may be used as a cue to initiate and coordinate sequences.

8.2.2 Human Gaze in HRC

The field of human-robot interaction often seeks inspiration from human-human interactions to create more natural and intuitive experiences. One research direction, inspired by this approach, explores the concept of robots mimicking human gaze behavior [Boucher et al., 2012; Moon et al., 2014; Stanton and Stevens, 2017; Hayashi and Mizuuchi, 2017; Faibish et al., 2022]. However, this typically requires robots to possess eye-like features and the ability to convey subtle cues like gaze direction and blinks. In industrial settings, cobots often lack such features, and attempts to anthropomorphize them often involve adding hardware like glasses or tablets [Fischer et al., 2015; Kühnlenz et al., 2020; Terzioğlu et al., 2020; Onnasch et al., 2023]. This chapter, however, focuses on understanding the patterns in human gaze behavior when collaborating with a robot.

To gain insights from the existing literature on how humans use gaze during collaborative tasks with robots, a literature survey was conducted using the Scopus database. However, to maintain the focus on the collaboration aspect, the review excluded studies (e.g,. Chadalavada et al. [2018]; Paul et al. [2023]; Weber et al. [2023]) that did not involve physical collaborative tasks, which encompass situations where humans and robots work together physically (e.g., assembling an object). Additionally, studies [Paletta et al., 2019; Upasani et al., 2023; Galvani et al., 2023] using gaze information to detect aspects like stress or task load were excluded as these aspects are not specific to collaboration.

A crucial aspect to consider in HRC studies is the nature of the shared task, particularly whether it is cooperative or collaborative. Several classifications and definitions exist for collaboration and cooperation [Kolbeinsson et al., 2019; Onnasch and Roesler, 2021]. For clarity, this chapter adopts a distinction commonly used in the HRC literature [Schmidtler et al., 2015; Bütepage and Kragic, 2017; Simões et al., 2022]. Cooperation is defined as the setting where the robot and the human work in a shared workspace at the same time, but on physically separate sub-tasks. Collaboration, on the other hand, involves joint manipulation of the same object, in addition to the aspects of cooperation (shared workspace and simultaneous work). While these terms are often used interchangeably, it is important to differentiate them because not all tasks with shared goals require the same level of interdependence between humans and robots. This distinction can lead to differences in how people use gaze during interaction. Tasks where humans and robots work on completely separate sub-tasks might not involve any gaze cues, unlike scenarios requiring frequent interaction between the human and robot. For example, an object handover task requires fine-grained coordination regarding the timing and location of handover [Strabala et al., 2013; Moon et al., 2014], which can be achieved through gaze cues. On the other hand, a task designed in the style of a conveyor belt - where one partner places the object on one side and the other partner picks it up - may not require either of the partners to perceive the other's gaze cues.

In addition to the nature of shared tasks, two aspects relating to gaze-based interaction - implicit/explicit and gaze requirement - shape gaze behaviors during HRC sessions. According to Ju [2015], implicit human-agent interactions occur when the agent possesses some level of agency and performs actions without explicit communication or knowledge of the user. In the context of gaze-based HRC interactions, implicit interactions may involve the robot monitoring and responding to the user's gaze behavior. However, the user is unaware of the gaze trigger for specific robot actions. On the other hand, explicit interaction studies inform the participants about the underlying gaze trigger mechanism and they communicate with the robot using these pre-defined gaze signals. For example, a pick-and-place scenario can be designed to utilize implicit or explicit gaze of the operator. In this case, implicit gaze interaction may involve the operator verbally communicating the selection but the robot monitors the operator's gaze to anticipate the choice (e.g., Huang and Mutlu [2016]). The same task can be designed to be explicit and solely gaze-based interaction if the operator is informed that the selection can be made by looking at the desired object for a few consecutive seconds (e.g., Shi et al. [2019]).

Regarding the aspect of gaze requirement, some studies require the participants to exhibit certain gaze behaviors to complete the task, whereas others utilize gaze to improve the interaction (e.g., reducing reaction time). In the previous pick-and-pack example, the implicit interaction does not have a gaze requirement as the task can be completed through verbal communication. However, the explicit interaction would inevitably require the operator to look at an object to complete the task. Although there is a huge overlap between explicit interactions and gaze requirements, implicit interactions can also involve gaze requirements. For instance, the study by Palinko et al. [2016] (described later) required the participants to exhibit a pre-defined gaze sequence to complete the task but they were not informed about the gaze sequence. The idea behind this type of interaction design is that the expected gaze behaviors are intuitive and occur naturally during the interaction, and hence, all the participants could complete the task even if they are unaware of the requirement.

Mehlmann et al. [2014] investigated gaze behavior in a cooperative puzzle game involving a humanoid robot (Nao) assisting a human participant in sorting puzzle pieces. To instruct the participant on where to move the pieces, the robot used a combination of gaze, speech, and pointing gestures. The gaze interactions were implicit and participants were not required to exhibit any specific gaze pattern. While their primary focus was the robot's social and referential gaze, they observed specific patterns in participant gaze as well. Notably, participants frequently employed mutual gaze (looking directly at the robot's face) to signal the robot that they were ending their turn. This finding highlights the potential role of gaze in turn-taking during robot-assisted tasks.

To explore implicit gaze-based communication, Palinko et al. [2016] devised a collaborative tower-building task with a humanoid robot (iCub). The robot held a block in each hand, and participants needed to trigger the robot to offer a block. To accomplish this, the participants had to look at the robot's face and then a specific hand (or vice versa). Importantly, participants were unaware of this trigger mechanism. They attempted various communication methods, including speech, gaze, and pointing. Notably, participants achieved an average success rate of 95%, with unsuccessful attempts attributed to failures in gaze detection. Despite being unaware of the specific trigger, participants achieved high success rates, suggesting that gaze is a natural and intuitive mode of communication.

A study by Ivaldi et al. [2017] investigated how personal characteristics like negative attitudes towards robots influence gaze patterns during collaborative tasks. In their study, participants worked with a humanoid robot (iCub) to assemble two cylindrical rolls. The participants first instructed the robot to hold the rolls, then guided its hands for proper alignment, and finally completed the assembly by taping them together while the robot held the rolls steady. Their results revealed correlations between negative attitudes towards robots and gaze behavior. Participants with higher negative attitudes exhibited a tendency to spend more time looking at the robot's hands and less time looking at its face. The study investigated implicit gaze behaviors with respect to personal characteristics and did not require gaze behaviors to complete the task.

Building upon the work of Mehlmann et al., Kurylo and Wilson [2019] analyzed gaze patterns during a collaborative medication sorting task involving a humanoid robot (Nao). Similar to the previous study, participants communicated with the robot using speech, gaze, and pointing gestures. Since their analysis was conducted on video recordings, the interactions were implicit and there were no pre-defined gaze requirements. Their research identified four distinct gaze patterns that emerged throughout the task: mutual gaze, confirmatory gaze, referential gaze, and looking away. Notably, they found that mutual gaze was often used when participants required assistance, similar to human-human interactions. Their findings provide further evidence that humans extend their natural gaze cues to interact effectively with humanoid robots.

In their study, Oliveira et al. [2018] investigated gaze behavior in a 4-player cooperative card game played in teams of two. Two humans and two robots with facial features (EMYS) participated in three sessions with rotating pairings. One robot was programmed to be supportive, while the other was competitive. The authors observed the gaze behaviors of the participants to derive insights about the implicit gaze behaviors during the game and the study did not mandate any gaze patterns. Among all partner configurations, the participants looked at the supportive robot more often than the competitive robot or the other human player. On the contrary, among all opponent configurations, the participants looked at the competitive robot more often than the supportive robot or human players. Their findings suggest that humans extend their natural tendency in social interactions to robots fulfilling similar roles.

Recent research has begun to explore interactions with non-humanoid mechanical robots, which are typically robotic arms. This distinction is crucial, as the absence of human-like features, such as faces, can potentially influence human behavior during interaction.

For example, Oka and Uchino [2016] explored cooperative conveyance using a specialized mechanical robot (LIEN) capable of performing nine object manipulation actions. In this study, the participants communicated their desired actions through voice commands while facing the robot, then looked away while the action was executed. The participants were explicitly informed on how to control the robot through gaze and speech (both signals were required). The participants successfully completed 95% of the tasks, with two low-margin failures (robot activated slightly late). Based on this performance and participants' feedback, the authors concluded that their proposed strategy, leveraging the combination of speech and gaze, was both effective and easy to learn.

Huang and Mutlu [2016] investigated the use of gaze information to enhance efficiency in HRC settings. They designed a cooperative smoothie-making scenario where a cobot (Kinova MICO robot arm) picked blocks representing fruits selected by the participant acting as the customer. While participants verbally communicated their choices, the study also monitored their gaze behavior. This gaze data was used by the cobot to anticipate the participant's selection, allowing it to grasp the chosen fruit block faster. As mentioned before, this study represents implicit gaze interaction, without any mandatory gaze patterns. However, it is important to note that the overall interaction cannot be considered implicit as the participants were instructed to communicate their choice through verbal commands.

Extending the research by Huang and Mutlu, Shi et al. [2019] employed a cobot (UR10) to pick up various office supplies lying on a table. In this study, the selection was communicated solely through the gaze of the participant. Similarly, Scalera et al. [2021] demonstrated gaze-based control of a cobot (UR5) for teleoperated artistic drawing tasks. These studies showcase the potential of gaze-based interfaces for controlling cobot actions. Both studies informed the participants on the gaze-based control (explicit interaction) and involved mandatory gaze behaviors for successful interaction.

Newman et al. [2020] further explored the potential of gaze for anticipatory robot actions in a handover game. In each round, the participant selected an item from three options, which the cobot (Kinova MICO) then picked and handed over. While the selection was communicated through button presses, they monitored participants' natural gaze behaviors to generate implicit anticipatory movement. Their findings suggest that gaze data collected shortly before the button press was most the most reliable indicator of participant's choices, compared to data collected throughout the entire round.

Table 8.1 summarizes the key information from the previously mentioned studies. Notably, the table reveals a predominance of cooperative tasks compared to collaborative tasks in the analyzed research. Furthermore, the table highlights an interesting trend: studies analyzing implicit human gaze tend to utilize robots with human-like features.

8.2.3 Research Gaps

This thesis focuses on industrial HRC scenarios which typically deploy cobots. As reflected in Table 8.1, existing research on gaze behavior in HRC settings involving cobots predominantly focuses on cooperative tasks rather than collaborative tasks involving joint object manipulation. This observation forms the foundation for this chapter, which aims to address the following research gaps:

- **Gaze-based Joint Activity Initiations:** While existing research analyzing human

	Paper	Task	Task type	Gaze Interaction
SOCIAL ROBOT	Mehlmann et al. [2014]	Sorting	Cooperative	I, NR
	Palinko et al. [2016]	Assembly	Cooperative	I, R
	Ivaldi et al. [2017]	Assembly	Collaborative	I, NR
	Kurylo and Wilson [2019]	Sorting	Cooperative	I, NR
	Oliveira et al. [2018]	Game	Cooperative	I, NR
COBOT	Oka and Uchino [2016]	Conveyance	Cooperative	E, R
	Huang and Mutlu [2016]	Pick&place	Cooperative	I, NR
	Shi et al. [2019]	Pick&place	Cooperative	E, R
	Scalera et al. [2021]	Drawing	Cooperative	E, R
	Newman et al. [2020]	Pick&place	Cooperative	I, NR
	Prajod et al. [2023b]*	Assembly	**Collaborative**	**I, NR**
	Lavit Nicora et al. [2024]*	Assembly	**Collaborative**	**I, R**

*Table 8.1: An overview of the literature studying human gaze in HRC settings. 'I' stands for implicit interaction and 'E' for explicit interaction. 'R' indicates that gaze was required for interaction and 'NR' indicates gaze was not required for completing the task. The entry marked with * is expanded in the subsequent sections of this chapter.*

gaze in HRC primarily focuses on cooperative tasks, it's valuable to consider insights from studies exploring robot's gaze in collaborative tasks, such as object handover. These tasks, often inspired by human-human interaction, involve joint object manipulation and investigate aspects like initiating joint activity [Strabala et al., 2013; Moon et al., 2014]. When humans collaborate with robots with facial features (eyes or a face), their gaze behavior may resemble human-human interactions. However, the question remains: what gaze patterns emerge when the robot lacks such features? Specifically, how might the participants initiate joint activities, and could gaze play a role in their strategies? To determine if this potential gaze behavior is natural, it's crucial to investigate it in settings that do not require specific behaviors for task completion. This allows for observation of natural gaze patterns and their potential use in initiating collaboration with robots lacking human-like features.

- **Automatic gaze-based cobot triggering**: Studies presented in Section 8.2.2 that require gaze for controlling the robot's actions often explicitly inform participants about the control mechanism. This approach prioritizes demonstrating the feasibility and usability of the system. An exception is the work by Palinko et al., who aimed to demonstrate the naturalness and intuitiveness of their system by allowing the participants discover the control mechanism on their own. However, they employed a humanoid robot with clear facial features. Whether a natural and intuitive gaze-based control system is feasible when collaborating with cobots lacking facial characteristics remains an open question.

8.3 Analyzing Gaze Behavior in HRC

The importance of gaze as a social cue in human-human collaboration is well-established, with mutual gaze and joint attention frequently observed in such interactions [Green et al., 2008; Pfeiffer et al., 2013; Admoni and Scassellati, 2017; Cañigueral and Hamilton, 2019; D'Angelo and Schneider, 2021]. Many studies have explored introducing gaze behavior in humanoid robots to improve collaboration [Srinivasan and Murphy, 2011; Ruhland et al., 2015; Admoni and Scassellati, 2017; Ajoudani et al., 2018]. However, in industrial HRC settings, cobots are typically robotic arms lacking human-like features for conveying social cues. While some studies have investigated anthropomorphizing cobots (e.g., Fischer et al. [2015]; Kühnlenz et al. [2020]; Terzioğlu et al. [2020]; Onnasch et al. [2023]), this chapter examines whether certain gaze behaviors observed in human-human interaction also manifest in HRC, even in the absence of added human-like characteristics.

Gaze-based social cues are essential for facilitating collaboration in human-human interactions. These cues, such as making eye contact and looking in a certain direction, play a key role in conveying intention and coordinating actions. For instance, participants may look at their collaborating partner to signal readiness or intend to collaborate. However, it is not known whether humans exhibit the same gaze cues observed in human-human interaction when collaborating with a cobot that lacks human-like features. This section aims to address this research question by analyzing the natural gaze behavior of participants working with a cobot on a collaborative assembly task.

8.3.1 Assembly Task

This analysis leverages the video data collected in Chapter 7. As described in the previous chapter, the HRC assembly task involved a participant and a cobot working together to build a 3D-printed planetary gearbox. The assembly process consisted of two distinct phases:

1. **Individual Assembly Phase**: Participant gathered components from the nearby box and produced a sub-assembly following the assembly steps. Meanwhile, the cobot performed a scanning motion over the pre-assembled sub-assemblies.

2. **Joint Activity**: The cobot delivered a pre-assembled sub-assembly to a designated collaboration area. After this, the cobot and participant collaborated by meshing the gears of their respective sub-assemblies, completing the gearbox.

8.3.2 Setup and Dataset

This analysis focuses on the “Adaptive" condition from Chapter 7, a “Wizard of Oz" setup where the joint activity occurred based on the participant's production rate. The experimenter (acting as the “wizard") triggered the cobot when the participant neared assembly completion, ensuring synchronized production rates. Since the participants were unaware of the trigger mechanism, they exhibited natural behavior and social cues, assuming the cobot was capable of synchronized collaboration. Importantly, the wizard controlled the cobot using sub-assembly completion information, not participant gaze. Therefore, participants were not required to exhibit specific gaze patterns for task completion. Video recordings from this condition were analyzed to identify participants' gaze patterns during collaborative assembly, particularly how they initiated the joint activity.

Figure 7.3 from Chapter 7 depicts the experimental setup. The experimental layout comprised three distinct areas: two individual workstations for the cobot and the operator to independently assemble their sub-assemblies, a shared workspace for collaborative joining of sub-assemblies (joint activity), and the wizard's workstation. The wizard's table is positioned opposite the cobot's workspace. This layout allows clear identification and distinction of the participant's gaze directed towards the wizard, their assembly table, or the cobot itself. Additionally, the wizard had a clear view of the participant's activity, ensuring timely cobot triggers.

This analysis utilized video recordings captured by frontal cameras from 37 participants. Each video lasted approximately 15 minutes, resulting in a total of 555 minutes of video material. These recordings captured the participants' upper body and face, allowing for analysis of their gaze direction. The cobot's wrist was occasionally visible in the videos, primarily during brief movements towards or away from the participant and during the joint activity phase of the production cycle. An example from the video samples is presented in Figure 8.2.

8.3.3 Annotations

The analysis focuses on whether participants utilize specific gaze behaviors to initiate the joint activity phase with the cobot. Two key pieces of information were required for this analysis: participant gaze direction and the start of joint activity.

Gaze Annotations

This analysis focuses on gaze areas rather than precise gaze estimation. Three key areas were identified within the environment: the cobot, the participant table (while assembling), and other locations (clock, window, etc.). Consequently, the annotation scheme employed three labels: 1 for gaze at table, 2 for gaze at cobot, and 0 for other directions.

Gaze direction labels were obtained through an attention recognition model described in Chapter 3. This model, trained using transfer learning, maps gaze estimations to designated areas of interest. Given face images as input, the model classifies gaze direction into one of the three labels. This approach significantly reduces the manual labeling efforts involved in annotating the entire video.

The training process utilized a transfer learning technique, leveraging the weights of an existing gaze estimation model. First, a convolutional neural network (VGG16 architecture) was trained on the ETH-XGaze face image dataset [Zhang et al., 2020] to estimate gaze direction in terms of pitch and yaw. Subsequently, the model's prediction layers were fine-tuned to map gaze onto the three defined areas of interest. Fine-tuning involved collecting volunteer images in a guided gaze setting mirroring the current study's setup. This process achieved an accuracy of 94.3% and an F1-score of 94%. To demonstrate robustness, the model was further validated in a non-guided setting. Additional details regarding training procedures and validation are available in Chapter 3.

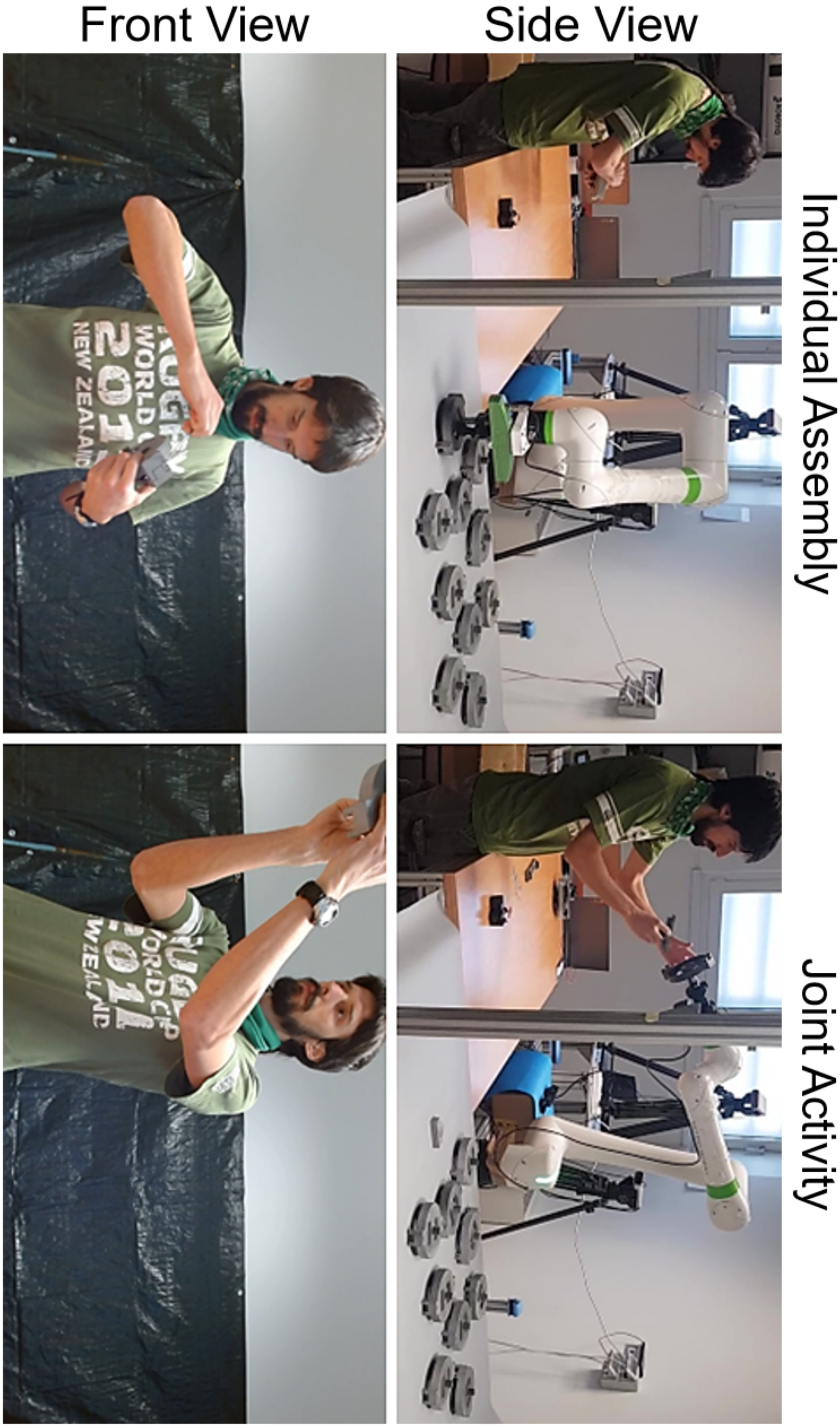


Figure 8.2: The front and side views of a participant during the individual assembly phase and joint activity of a production cycle. Only the front view is annotated for the gaze behavior analysis. The copyright remains with the authors [Prajod et al., 2023b].

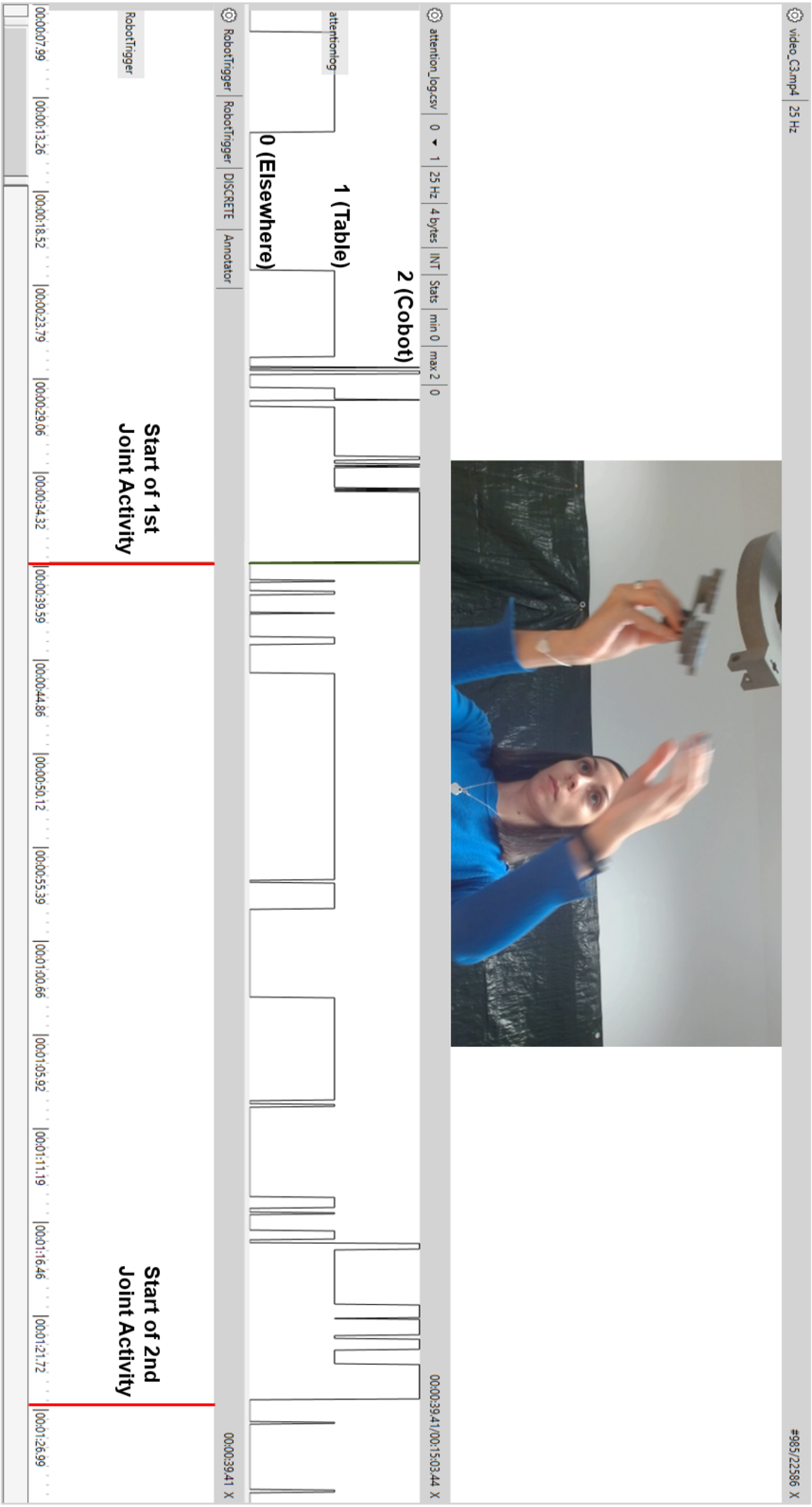


Figure 8.3: A snap of NOVA interface: gaze recognition predictions are displayed in the top track, while red lines in the bottom track mark the start of joint activities. The copyright remains with the authors [Prajod et al., 2023b].

Joint Activity Annotations

The participant activities involve: assembling their own sub-assembly and jointly meshing sub-assemblies with the cobot. This analysis focuses on the few seconds leading up to the joint activity. Therefore, the frame where the cobot reaches the participant (i.e., stops in front of them) for the joint activity was annotated for each assembly cycle. The participant gaze behavior in the few seconds preceding this point was analyzed.

The video annotation process utilized the NOVA tool [Baur et al., 2013]. This tool facilitated not only annotating the relevant frames but also visualizing the predictions from the attention recognition model as an annotation stream. A total of 585 joint activities were labeled, with an average of 15.8 activities per participant.

8.3.4 Analysis Procedure

Visual Inspection

As a starting point, NOVA visualizations were utilized to examine participant gaze patterns relative to the joint activity start. An example visualization is presented in Figure 8.3. The bottom track displays the annotated joint activity starting points, while the top track shows the predicted gaze annotations with values of 0, 1, or 2 corresponding to different gaze directions.

The analysis specifically focuses on instances where the predicted class is 2, indicating that the gaze was directed towards the cobot. A promising trend is observed in the top track, with spikes (class = 2) appearing in the few seconds preceding the joint activity. This pattern suggests that participants might be looking at the cobot to potentially initiate the collaborative phase.

Quantitative Analysis

This part of the analysis aimed to quantify how often participants used gaze towards the cobot to plausibly initiate joint activities and to distinguish these intentional gazes from occurrences unrelated to the joint activity phase. To this end, the following multi-step procedure was devised.

1. **Time window**: The analysis calculated the number of participant gazes towards the cobot within 15 seconds before each joint activity. This timeframe considered the cobot's movement time: 10-12 seconds to reach the part, grab it, and pick it up, and 3 seconds to move to the collaborative joining position.

2. **Cobot gaze instances**: Predictions from the attention recognition model were smoothed using a three-point moving window to reduce jitters in continuous predictions. A peak detection algorithm identified instances where the smoothed data indicated participant gaze towards the cobot. To ensure sustained gaze and minimize spurious detections, only peaks spanning at least five frames (at 25 fps) were considered, indicating the participant looked at the cobot for at least five consecutive frames.

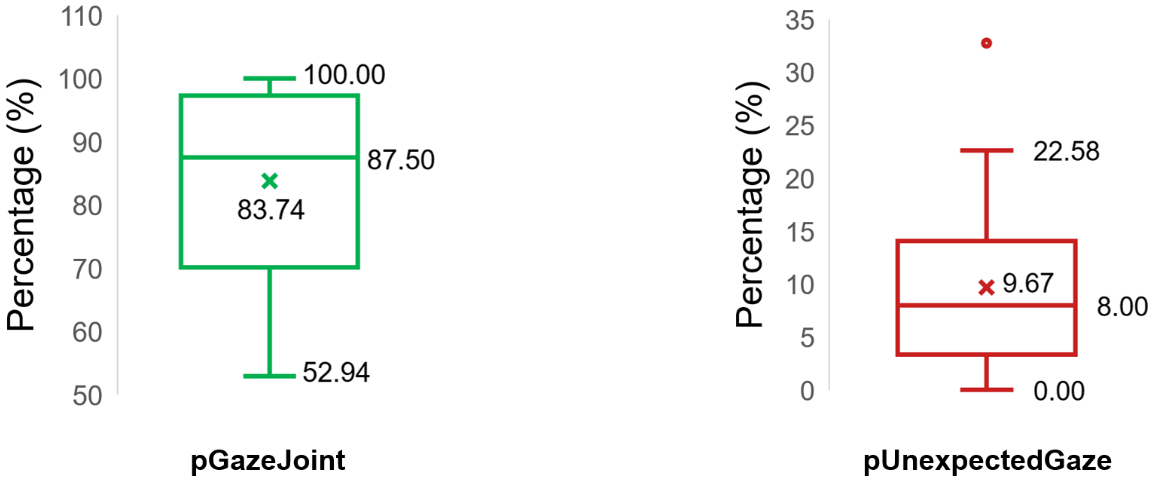


Figure 8.4: Box plots illustrating the distribution of pGazeJoint and pUnexpectedGaze values across all participants. The horizontal line in the middle of the box represents the median value, while the "X" symbol denotes the mean value for each distribution.

3. **Gaze-preceded joint activities (*pGazeJoint*)**: Using identified gaze peaks and annotated joint activity start points, the analysis calculated the percentage of joint activities preceded by participant gaze towards the cobot. A joint activity was deemed "gaze-preceded" if the participant looked at the cobot at least once within the 15-second window before its start. This metric (expressed as a percentage) represents the proportion of gaze-initiated joint activities compared to the total number of joint activities in a session.

4. **Unexpected gazes to cobot (*pUnexpectedGaze*)**: The analysis expected participants to gaze at the cobot for two primary reasons - initiating joint activity and during the activity itself, which typically lasted 20-25 seconds. Therefore, any gaze towards the cobot outside this timeframe is considered "unexpected" and potentially unrelated to the collaborative task. To quantify this unexpected gaze behavior, a metric called *pUnexpectedGaze* is calculated. This metric, expressed as a percentage, represents the ratio of unexpected gazes towards the cobot to the total number of gazes towards the cobot.

8.3.5 Analysis Results

Figure 8.4 presents boxplots displaying the distribution of *pGazeJoint* and *pUnexpectedGaze* values obtained from all the participants. The mean *pGazeJoint* value is 83.74%, indicating that, on average, 83.74% of collaborative joining activities were preceded by a gaze towards the cobot. This suggests a strong association between looking-at-cobot behavior and joint activity initiation.

Meanwhile, the mean *pUnexpectedGaze* is only 9.67%, signifying that very few gazes towards the cobot occurred outside the expected timeframe associated with the joint activity. This finding further supports the conclusion that looking-at-cobot behavior primarily occurs around the collaborative joining phase.

8.3.6 Insights

The analysis results revealed a tendency for participants to look at the cobot when they were ready for joint activity, evidenced by the high *pGazeJoint*. This behavior, reminiscent of human-human interaction, potentially serves as a social cue to initiate joint activity, promoting more natural and intuitive human-robot collaboration.

Furthermore, the analysis indicates that gaze directed towards the cobot typically occurs around the collaborative joining activity timeframe, supported by the low *pUnexpectedGaze* value. Interestingly, longer joining times were identified as a contributor to unexpected gazes. During some assembly cycles, participants took more time than anticipated to align sub-assemblies, leading to a collaborative joining process exceeding the estimated duration. Additionally, unexpected software behaviors or delays in the cobot's performance contributed to unexpected gazes. In some cases, the cobot did not immediately initiate the subsequent assembly cycle after completing the previous one, leading to a few unforeseen seconds of delay before the next cobot movement. This delay captured the participants' attention and prompted them to look towards the cobot to monitor the situation.

While not formally analyzed in the previous chapter, valuable insights emerged from participants' comments. All comments were originally in Italian and are presented here in translation. One of the participants (Participant 3) mentioned: "I noticed that the robot was synchronized with me and I thought it might be because of the camera, so I tried looking at it to see what would happen". Another participant (Participant 37) said: "In some cases, I was surprised by how slow the robot was, so I tried looking at it in the hope of making it faster". These participants believed their gaze influenced the cobot's behavior, when in reality, the wizard controlled it based on their sub-assembly completion. These comments highlight the intuitive nature of gaze as a communication tool and its potential role in collaborative interactions.

Moreover, Participant 15 suggested that "adding eyes" to the cobot could make it more expressive. This suggestion, while outside the scope of this work due to its focus on anthropomorphism, underscores the potential for gaze-based communication to enhance naturalness and intuitiveness in HRC.

8.4 Towards Gaze-based Triggers in HRC

Building on the gaze analysis in Section 8.3, this study aims to pilot a fully integrated augmented collaborative cell where joint actions are automatically triggered based on the participant's detected gaze behavior. The study objective extends beyond technical feasibility, investigating whether a cobot aware of gaze cues fosters a more natural and intuitive collaboration experience.

Inspired by Palinko et al. [2016], natural and intuitive collaboration is defined as the ability of participants to successfully complete the task without explicit instructions about the trigger mechanism. In other words, participants must discover how to initiate joint activity through their own interaction experience.

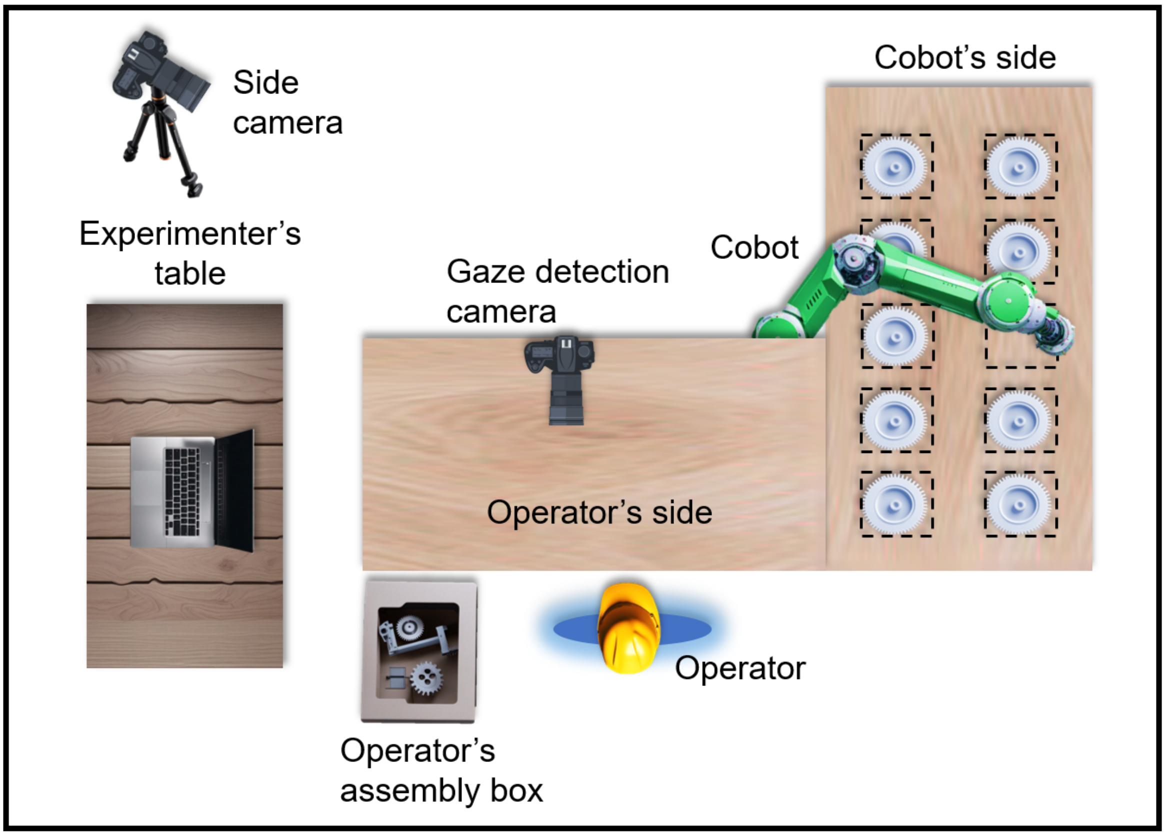


Figure 8.5: The layout of the experimental setup for automatic gaze-based cobot triggering system. The layout is same as Chapter 7, except the addition of the side camera (near the experimenter)

8.4.1 Experimental Setup

While Section 8.3 demonstrated a specific gaze behavior (looking at the cobot) preceding joint activities, it is crucial to acknowledge that the data was not originally collected for this chapter's specific research questions. Therefore, a dedicated study is necessary to assess the feasibility of an automatic gaze-based cobot triggering system.

The participant's task remained identical to Section 8.3.1. However, the cobot's behavior during the individual assembly phase differed slightly, as detailed in Section 8.4.2. Notably, due to the implementation of automatic gaze-based triggering (see Section 8.4.3), the "wizard" role became obsolete. Although absent in the triggering process, an experimenter remained present to address any potential technical issues.

For the subsequent analysis, a broader view of the setup including the participant and the cobot was required. Hence, an additional camera was installed near the experimenter's table, capturing the entire scene as shown in Figure 8.5. While the front camera facilitated automated triggers, the primary data for the analysis were video recordings from this side camera.

8.4.2 Cobot Operating Modes

Following the three experimental sessions described in Chapter 7, a few participants reported that the cobot's scanning motion during the individual assembly phase was noisy and distracting. To investigate if this motion influenced participant behavior, two experimental conditions were implemented:

- **Scanning**: This condition replicated the previous cobot behavior (Section 8.3.1), where the cobot scanned over the pre-assembled parts during the individual assembly phase while waiting for the trigger. However, unlike the previous setup, the trigger was now automatically generated based on the participant's gaze behavior.
- **Still**: In this condition, the cobot remained stationary above the pre-assembled components instead of performing the scanning motion. Upon receiving the gaze-based trigger, the cobot moved to a specific pre-assembly, picked it, and brought it to the participant for the joint activity.

These two conditions were designed to determine if the scanning motion influenced the participants' behavior. Additionally, the Still condition is expected to potentially facilitate the participant's understanding of the gaze-based trigger mechanism. Since the cobot wouldn't initiate any action until the participant looked at it, this condition might provide clearer cues about the role of gaze in initiating the joint activity.

8.4.3 Real-time Implementation

To automatically trigger the cobot in real-time, this study utilizes two sub-systems: automatic gaze recognition and generating cobot triggers. The overall cobot behavior, including movement and interaction control, was implemented using the Robot Operating System (ROS) Noetic [Quigley et al., 2009].

Gaze Detection

This sub-system leveraged the same attention recognition model employed for automatic annotations in Section 8.3.3 to detect participant gaze direction in real-time. It was implemented as a pipeline within the SSI framework [Wagner et al., 2013], a Windows-based platform designed specifically for recording, processing, and analyzing social signals. The sub-system operated in four steps:

1. Input: Upper-body video captured by the front camera served as the initial input for the sub-system. Each frame of the video was processed separately in the pipeline.
2. Face detection: MediaPipe's face detection model Bazarevsky et al. [2019] was used to crop the input frame and focus solely on the participant's face region. The cropped facial images were scaled to 224 × 224 pixels (default VGG16 dimensions). As discussed in Chapter 6, participants occasionally exit the camera's field of view to retrieve additional boxes or bend down to pick up pieces. In these instances, face detection fails and the missing frames are replaced by a default, solid-colored image.

3. Gaze classification: The cropped and scaled face images were fed into the previously mentioned attention recognition model, allowing for real-time classification of the participant's gaze direction (cobot, table, elsewhere). Upon receiving a solid-colored image (indicating face detection failure) from the face detection module, this module classifies the individual's gaze direction as "elsewhere".

4. Output: The classification results for each frame were transmitted to the cobot-triggering sub-system via UDP sockets for further processing.

Cobot Triggers

The cobot triggering sub-system utilizes the VSM framework [Gebhard et al., 2012] to execute the task logic designed for the experiment. VSM communicates with the ROS master (communication hub) through topics and services, allowing for external control of the cobot.

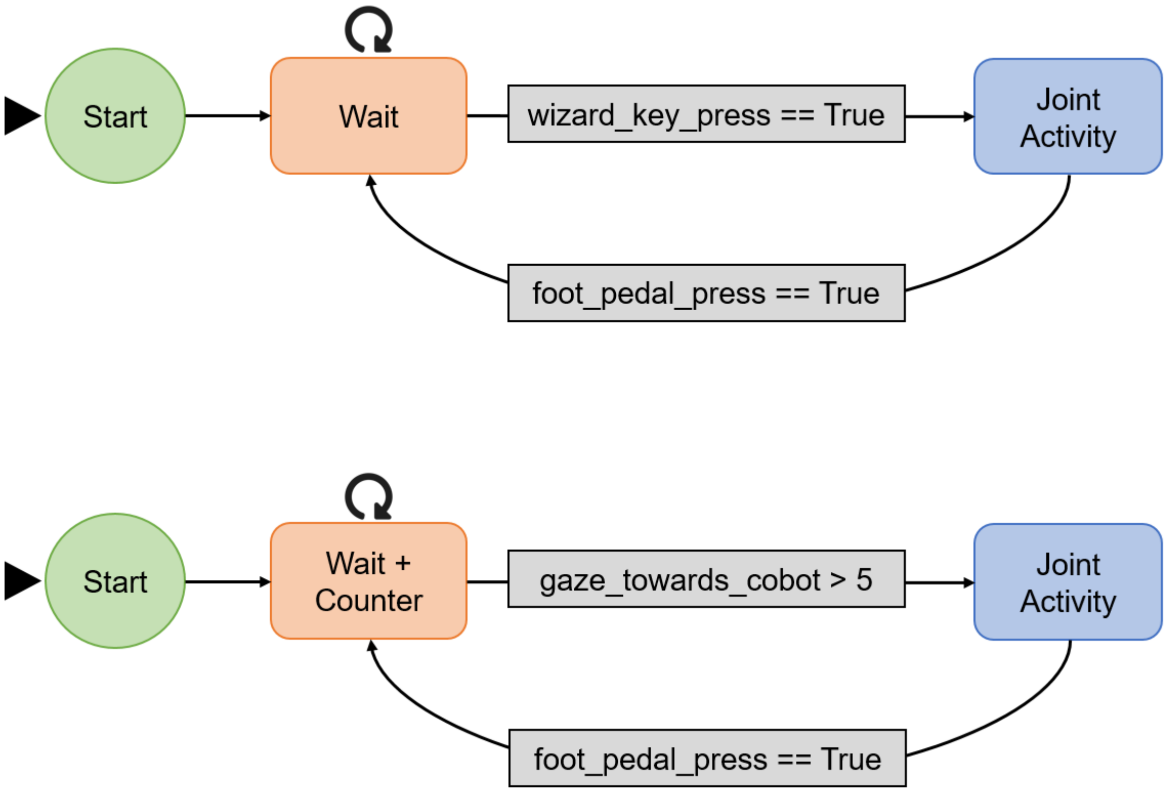


Figure 8.6: An illustration of the cobot triggering sub-system in both the previous Wizard-of-Oz setting (top) and Real-time implementation (bottom). In the "wait" state, the cobot performs the individual assembly actions (scanning or still) until it receives the trigger signal. Upon receiving the trigger, the cobot enters the "joint activity" state until the participant presses the foot pedal. The action returns the cobot to the "wait" state for the next production cycle.

In the previously utilized "Adaptive" condition (see Section 8.3.2), VSM listened for a specific keyboard press before triggering the cobot's movement for the joint activity (refer to Figure 8.6 for a simplified illustration). However, for the current study, the triggering logic was modified to rely on the participant's gaze data received from the automatic gaze recognition sub-system.

The VSM program received the gaze classification data from the SSI framework. In line with the analysis presented in Section 8.3.4, a valid trigger was generated only if the participant's gaze was detected towards the cobot for more than five consecutive frames. To achieve this, a counter was implemented in VSM to track the number of consecutive frames where the participant's gaze was directed towards the cobot. If the gaze direction changed, the counter was reset. The trigger for joint activity was sent to ROS only when the counter exceeded the predetermined threshold of five consecutive frames.

8.4.4 Data Collection and Annotations

Participants

A total of 10 volunteers participated in the current study. The group was demographically balanced with 5 male and 5 female participants, and an age range of 18 to 30 years (mean = 23.8, SD = 5.14). All participants were Italian and predominantly students from a university near the National Research Council of Italy - Lecco campus.

Importantly, the study included one participant with high-functioning ASD, while the remaining nine participants were neurotypical. This inclusion aimed to explore the feasibility of the gaze-based cobot system outside the behavioral patterns established by the entirely neurotypical group analyzed in Section 8.3. Previous research (Chapter 6) suggests differences in gaze behavior between neurotypical individuals and those with ASD during collaborative assembly tasks.

Study Protocol

Participants were initially informed about data treatment procedures and provided signed consent forms. They then underwent a brief training session to practice assembling gearboxes. Importantly, none of the participants had prior experience with the cobot, and they were not informed about the gaze-based automatic triggering system.

The study employed a within-group experimental design (Figure 8.7). Each participant interacted with the cobot under both the Scanning and "Still" conditions. Each condition lasted for the time required to assemble 10 complete gearboxes, with a short break between sessions. The order of the conditions was randomized and counterbalanced to control for any potential effects of experiencing one condition before another.

Following the second experimental session, participants were asked to share their impressions of the system. Their responses were recorded in Italian and later translated into English. Finally, the participants were debriefed about the automatic gaze-based triggering system and the overall goals of the study. This part of the study was also covered by the ethical approval from Commissione per l'Etica e l'Integrità nella Ricerca of the National Research Council of Italy (protocol n. 0085720/2022 of 23/11/2022)

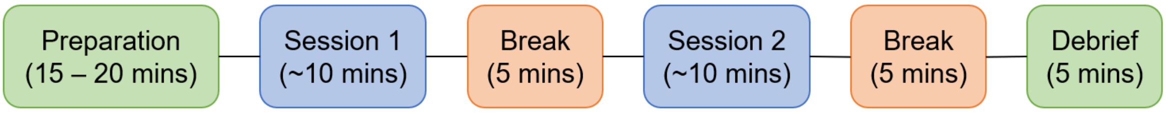


Figure 8.7: An overview of the experimental protocol consisting of two conditions (Scanning, Still). Each participant completed both conditions, but the order was alternated and counterbalanced across participants.

Annotations

In this study, a production cycle is said to be "successful" if the participant triggers the cobot for joint activity at the appropriate time (right before or immediately after completing their sub-assembly) and within a reasonable timeframe. Specifically, the trigger must be initiated within a maximum of 5 seconds after completing their sub-assembly, a threshold inspired by the work of Eldardeer et al. [2021].

To track the interaction timings and analyze the trigger dynamics, the following frames were annotated:

- **Assembly Completion**: The moment the participant finishes their individual sub-assembly.

- **Cobot Trigger**: The moment the cobot receives the trigger and begins moving towards its sub-assembly.

8.4.5 Analysis

Successful Initiations

The initial production cycle in each condition was excluded due to the potential influence of the researcher's start signal. Analyzing successful interactions (triggered within 5 seconds), the system achieved an overall success rate of 91.53% (88.64% for Scanning, 94.38% for Still).

While participants looked at the cobot and triggered joint activity in every cycle, some interactions exceeding the 5-second threshold were not classified as successful. Notably, the success rate in both conditions exceeded the gaze-preceded joint activities (*pGazeJoint*) observed in Section 8.3.4 (83.74%), indicating successful system implementation.

Waiting Time

Figure 8.8 presents the average waiting times of each participant in the Scanning and Still conditions. The green circle on Participant 8 indicates that the participant was characterized by ASD. In the Scanning condition, participants waited an average of 3.63 seconds after finishing their part to trigger the cobot. Notably, the Still condition led to shorter average waiting times of 2.73 seconds, likely due to not needing to wait for the ongoing scanning motion to complete.

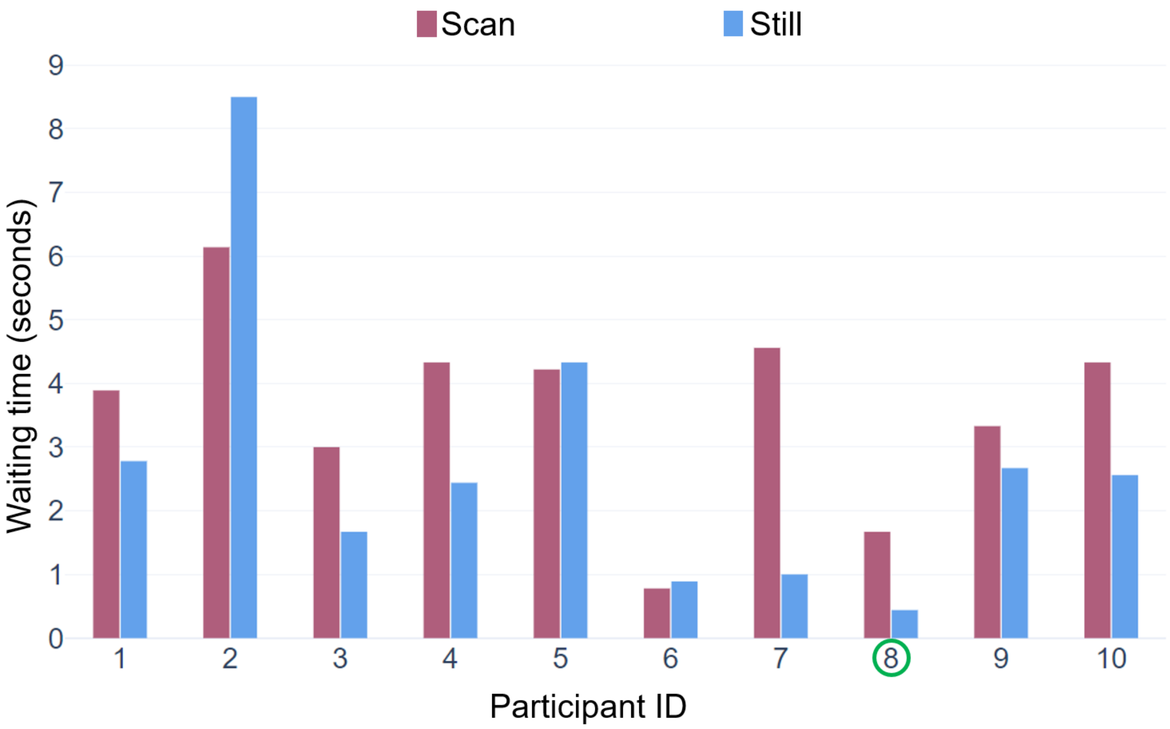


Figure 8.8: Bar graphs showing the average duration (in seconds) for which each participant waited for the joint activity to begin. For each participant, the first bar (magenta) represents the waiting time in the Scanning condition, while the second bar (blue) represents the waiting time in the Still condition. Participant 8, highlighted by a green circle, has been diagnosed with ASD.

Early Activations

Figure 8.9 shows the average duration of early activations in both conditions, i.e., instances where the cobot received the trigger before the participant finished their assembly task. The green circle indicates the average for the ASD participant. Overall, this occurred in 19.21% of interactions, with an average early-trigger time of 2.19 seconds.

This observation suggests that some participants may have learned the role of gaze over time, and began looking at the cobot before finishing to reduce waiting times. This hypothesis is supported by the higher average percentage of early activations observed among participants who reported understanding the gaze-based mechanism (58.82%) compared to those who did not. Excluding the ASD participant, around 90.91% of the early activation instances were triggered by the participants who commented that their gaze influenced the cobot's timing.

ASD Participant Behavior

Interestingly, the ASD participant exhibited a unique gaze pattern compared to the rest of the group. Unlike others, they often looked towards the cobot and triggered it before starting a new sub-assembly, as shown in Figure 8.10. This behavior resulted in a considerably high average early-trigger time of 15.50 seconds compared to the rest of the group.

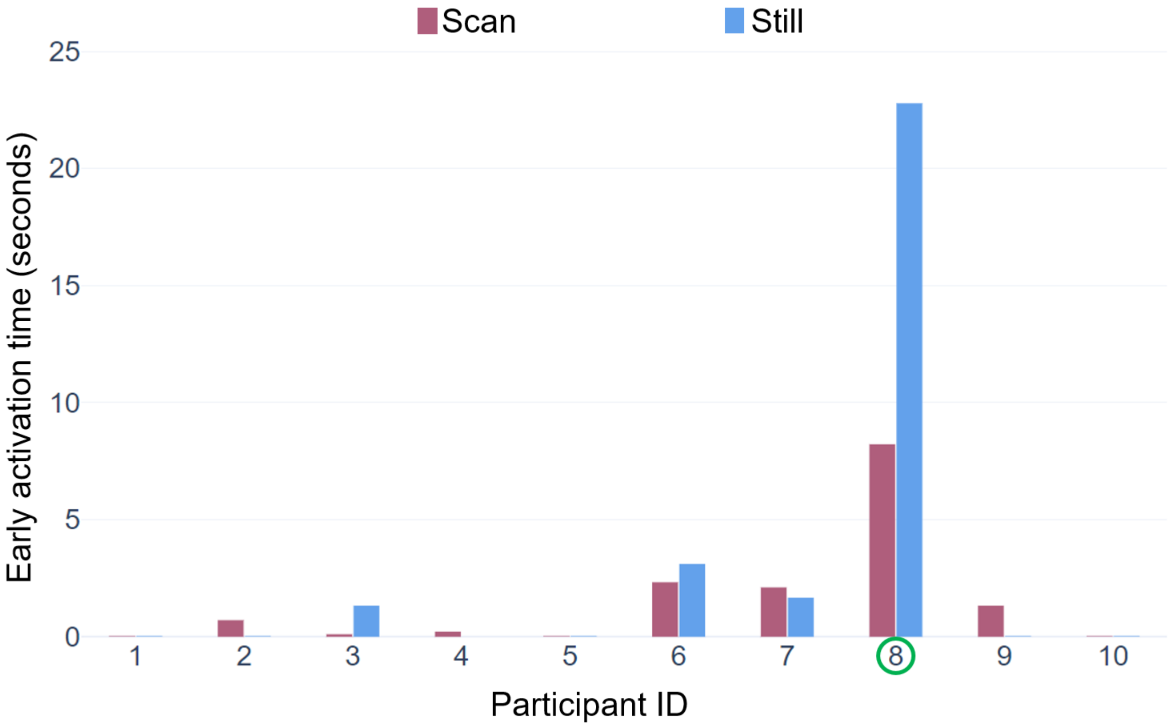


Figure 8.9: Bar graphs visualizing the average duration (in seconds) of early cobot triggers by each participant. Two bars represent the data of each participant: the first bar (magenta) represents the Scanning condition, while the second bar (blue) represents the Still condition. Participant 8, diagnosed with ASD, is highlighted with a green circle.

8.4.6 Insights

The fully integrated gaze-based cobot triggering system achieved a success rate of 91.53%, demonstrating the viability of using participants' gaze as a natural cue for triggering collaborative activities with a cobot. Most participants reported a positive collaboration experience, supporting the hypothesis that leveraging natural gaze behavior can enhance human-robot collaboration. However, Participant 1 commented: "The noise and the waiting times of the robot were irritating", highlighting the importance of considering all environmental factors for optimal working conditions.

As anticipated, most participants recognized that some aspect of their actions triggered the cobot's movement. This awareness was further supported by the observed early activations. Participants often reported feeling a greater sense of control over the system in the Still condition compared to the Scanning condition. For instance, Participant 2 said: "I think that during the scanning session, the robot had a fixed time before coming towards me. While in the still session, it came when I was done with my part". This suggests that the scanning motion may have masked the immediate responsiveness of the cobot, making the cause-and-effect relationship between their actions and the cobot's actions less obvious. Notably, four out of ten participants (40%) accurately identified their gaze behavior as the triggering mechanism. The remaining participants attributed the cobot's response to either a pre-defined schedule or alternative factors, such as their body position or the action of lifting the sub-assembly.

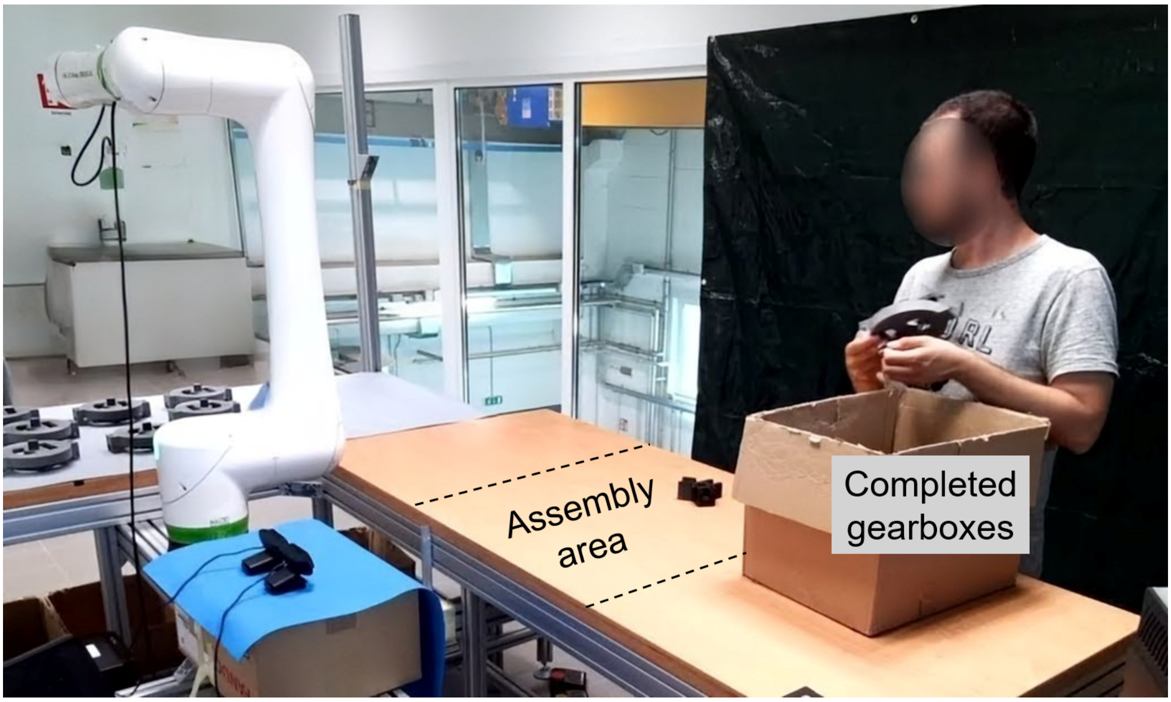


Figure 8.10: A participant with ASD triggering the cobot for joint activity through their gaze. The participant is about to drop the completed assembly into the box for completed assemblies. The area typically used for individual assembly is empty, showing that the participant triggered the cobot before assembling their part. The copyright remains with the authors [Lavit Nicora et al., 2024].

A majority of participants expressed a preference for the Still condition, where the cobot didn't perform a scanning motion. They felt the cobot reacted more quickly and was better synchronized with their actions. This preference aligns with the shorter average waiting times observed in the Still condition. In the Still condition, the cobot started movement towards the sub-assembly immediately upon receiving the gaze trigger. Conversely, in the Scanning condition, any trigger required the cobot to first interrupt its scanning motion before moving towards the sub-assembly, introducing a slight delay. Despite this difference in perceived responsiveness, participants successfully initiated joint activity with the cobot in both conditions. This observation indicates that while the scanning motion may have affected user preference, it did not influence the natural gaze behavior for triggering joint actions.

A unique pattern emerged when piloting the system with the participant diagnosed with ASD. The cobot was often triggered significantly earlier, resulting in the cobot waiting for the participant to complete the joint activity. This pattern was more pronounced in the Still condition. One possible explanation is that the cobot's stillness may have been unconsciously perceived by the participant as a potential malfunction, prompting them to monitor the cobot's status. This attention towards the cobot generated the trigger for the cobot's joint activity behavior, after which the participant resumed their task. Interestingly, the participant reported no perceived difference between the two conditions, stating, "I felt smooth working with the robot during both conditions". While this early triggering did not cause discomfort or hinder task completion, it underscores the importance of considering diverse needs and potential behavioral variations when designing human-robot collaboration systems.

8.5 Reflections and Remarks

This chapter investigated the potential of utilizing natural gaze patterns for a more natural and intuitive HRC in industrial settings, particularly when working with cobots that lack human-like features. This investigation was carried out in two complementary experiments.

Leveraging data from the previous chapter, the first experiment examined natural, unforced gaze behavior during an HRC task. Analysis revealed a distinct "look-at-cobot" gaze behavior employed by participants to initiate joint activity. Building upon this observation, the second experiment implemented a real-time gaze-based communication system. This system triggered the cobot to initiate joint activity based on the participants' gaze behavior. The gaze-based cobot-triggering system successfully achieved seamless collaboration with a very high success rate in initiating joint activity.

While participants reported positive collaboration experiences, the influence of the system on specific aspects like well-being and trust requires further investigation. Additionally, human activity recognition alongside gaze-based triggering could potentially further enhance the collaboration experience. By understanding the participant's progress with their task and tailoring the cobot's actions accordingly, the collaboration can become more seamless and efficient.

This chapter demonstrates the versatility of gaze information. The attention recognition model presented in Chapter 3, originally designed for distraction detection, was successfully repurposed here to initiate joint activity based on participants' attention towards the cobot. However, the study revealed an interesting difference in how one participant characterized by ASD interacted with the real-time system. This highlights the need for further research to explore how such gaze-based systems can be adapted to respond naturally to users with diverse needs and abilities.

Part IV

Conclusion

Chapter 9

Summary and Contributions

9.1 Key Takeaways

This thesis explored a range of worker states and developed machine learning models to predict these states. Building upon the foundation established in Part 1, the chapters in Part 2 investigated specific states: attention/distraction (Chapter 3), pain (Chapter 4), and stress (Chapter 5). While each chapter addressed specific research questions related to individual states based on existing literature gaps, this section revisits the overarching research questions from Chapter 1 regarding model applicability in real-world scenarios. Insights from each chapter are leveraged to answer these questions below.

The first overarching question explored the applicability of models across contexts: **Can models trained on datasets from different contexts be effectively applied to industrial Human-Robot Collaboration (HRC) settings? Or are these models specific to the training context?**

- **Attention Recognition**: The attention recognition models are not expected to be applicable to various settings as the areas of interest are tied to the work cell layout. Chapter 3 demonstrated that models trained on data from a specific layout with guided gaze can be applied to images from industry-like HRC sessions with the same layout. However, a limitation was identified in an assumption regarding distraction, which considered looking at non-assembly areas as a primary manifestation of distraction during down-time. It was observed that, during non-assembly periods participants sometimes looked at assembly components or even fidgeted them without actively assembling them. This observation highlights the need for models incorporating additional data like proximity to the assembling space and body pose for improved accuracy.

- **Pain Detection**: Chapter 4 showed that a model trained on one of the pain datasets demonstrated applicability to other contexts. Notably, although eye closure is often associated with pain expression, this dataset did not exclusively exhibit eye closure in pain images. This variation is realistic as eye closure is not always an indicator of pain (e.g., blinks). Consequently, pain predictions relied more heavily on other facial features like grimaces. These findings suggest that detecting pain in industrial settings using models trained on datasets from different contexts is a feasible goal, provided the chosen datasets have realistic variations.

- **Stress Detection**: The research presented in Chapter 5 suggests that Heart Rate Variability (HRV) models trained on datasets involving the same type of stressor demonstrate robustness across different contexts, even with differences in the intensity of experienced stress. So, for developing stress models applicable to industrial HRC settings, it is essential to select training datasets that match the type of stressor that occur frequently in such settings.

The second overarching question centered on assessing the learned features: **Are the features learned by the model generic and applicable to real-world scenarios, or are they specific to the limited data available?** This question is crucial when the size of the dataset is small.

- **Attention Recognition**: Due to the limited size of the dataset collected for a specific layout, transfer learning was employed to train deep neural networks for attention recognition. A large image-based gaze estimation dataset was used to train the source model. Here, only the final layers were fine-tuned to map gaze direction to an area of interest. This approach preserved the features learned for gaze estimation, making them independent of the target dataset. The effectiveness of these features in predicting attention in two different scenarios indicates their potential applicability across various situations.

- **Pain Detection**: Two pain detection models (trained on separate datasets) utilized transfer learning to leverage features learned by an emotion recognition model. The investigation of learned representations using Explainable Artificial Intelligence (XAI) techniques revealed that the models relied on patterns typical associated with pain expressions rather than being dataset-specific. However, the relative importance of these features for pain prediction varied across datasets.

- **Stress Detection**: The stress datasets were not very small compared to the attention and pain datasets. So, transfer learning techniques were not employed in training stress detection models. Deep learning models based on Electrocardiogram (ECG) achieved good performance on the training dataset but struggled to predict stress in a different dataset. This suggests that these models likely learned dataset-specific features because of overfitting. In contrast, HRV features exhibited better performance across datasets. Moreover, the HRV models demonstrated good cross-dataset performance on social stress datasets recorded in different contexts, indicating that the data was sufficient to train generic social stress detection models.

This section has thus far discussed the applicability of models trained for states that are relatively quick to manifest. The third overarching question addressed the long-term experiences: **What other relevant worker states might manifest during long-term industrial HRC?** To address this question, a week-long study was conducted in a lab work cell that mimicked an industrial HRC scenario. Participants with Autism Spectrum Disorder (ASD) were also included in this study to draw insights applicable to a wider population. By analyzing observations from the study through both quantitative and qualitative methods, the following patterns were identified and further explored:

- **Flow States**: Both neurotypical and ASD participants showed patterns of tiredness and boredom, with these manifestations increasing over time. Interestingly, the neurotypical participants exhibited a tendency to prioritize joint activity with the cobot over their ongoing tasks, seemingly to avoid delays for the cobot. Furthermore, they demonstrated adaptations in their workflow to better synchronize with the cobot's arrival for collaborative tasks. These behaviors were observed less frequently in the ASD participants. These patterns can be viewed from the perspective of the flow theory. In this framework, the cobot's production rate can be seen as the challenge level for the worker. When the cobot operates slower than the participant, waiting periods lead to boredom. Conversely, when the cobot is unexpectedly faster, participants adapt their activities to meet the increased challenge. For facilitating a flow experience, the cobot should ideally adapt to the worker's pace rather than the other way around. While the study did not identify clear manifestations of anxiety, further investigation into this area is warranted.

- **Gaze Cues**: Both neurotypical and ASD participants looked towards the cobot while waiting for it to complete its tasks. However, neurotypical participants tended to maintain gaze contact with the cobot for longer durations. Conversely, gazes towards the cobot were minimal while participants were actively assembling components. These observations suggest that gaze patterns may provide valuable social cues that can be leveraged by the cobot as a communication modality.

Building on the insights from the long-term study, a follow-up question was: **Can machine learning models be applied to detect these newly identified states?** To address this question, data collection was necessary to capture the identified state manifestations. The cobot's behavior was adapted to elicit these manifestations within shorter experimental sessions. Beyond training new models for these states, Chapters 7 and 8 explored whether the models or features developed in Part 2 could be leveraged for detection purposes.

- **Flow States**: The previously trained emotion recognition model and HRV feature extraction techniques from the stress detection research were explored for their applicability in detecting flow states. The analysis revealed that HRV features exhibited promising capabilities in differentiating between boredom, anxiety, and flow states. These findings led to the development and training of flow detection models specifically using HRV features. This demonstrates the applicability of HRV features in detecting not only stress but also flow states.

- **Gaze Cues**: One specific gaze cue identified as potentially informative was the act of looking at the cobot. This cue could be effectively detected using the attention recognition model developed in Part 2, which classified gaze direction as towards the cobot, table, or indicating distraction. Analysis of gaze behavior revealed that participants primarily looked at the cobot when they were ready for the joint activity.

This suggests that gaze cues can serve as triggers for initiating collaborative actions between the operator and the cobot. A system leveraging the attention recognition model to automatically trigger the cobot for joint activity was developed and tested. All participants successfully collaborated with this system. Interestingly, the sole participant with ASD consistently triggered the cobot much earlier than just-in-time for the joint activity, deviating from the behavior of the neurotypical participants. This suggests that although both groups exhibited similar gaze cues, the underlying intentions behind these behaviors might differ.

9.2 Contributions

This thesis addressed various questions pertaining to the applicability of specific machine learning models to industrial HRC settings. The contributions of this thesis include data collection, development and deployment of models, and insights about the relevant worker states. These insights is broadly classified as analytical or demonstrative, based on the type of presented observations. This section presents a brief account of the various contributions of this thesis.

9.2.1 Data Acquisitions

While some of the models were trained on publicly available datasets, others relied on data collected in specialized scenarios presented in this thesis. Excluding very small datasets collected for fine-tuning a few models, there were primarily three datasets collected as part of this thesis:

- **Social Stress Dataset**: The need for a social stress dataset which does not use a standardized stress inducing test was outlined in Chapter 5. To address this need, the chapter presented a multimodal social stress dataset that utilized a simulated job interview scenario to elicit social stress.

- **Long-term Industry-like HRC Dataset**: A key need for assessing applicability stemmed from the lack of long-term studies in industrial HRC settings. Chapter 6 addressed this research gap by designing an HRC work cell in the lab, which mimicked an industrial scenario. This setup was used to collect video clips and other observational data intermittently from the week-long study. This setup facilitated the natural occurrence of states such as distractions and boredom. Moreover, it facilitated the analysis of behavioral patterns in neurotypical and ASD participants and how they evolve over time.

- **Flow in Industry-like HRC Dataset**: The insights from the long-term study hinted at the need for detecting flow-related states, plausibly induced by the production rate of the cobot. This led to the collection of a multimodal dataset that recorded the responses of participants during three cobot production rate conditions (Slow, Fast, and Adaptive).

9.2.2 Implementations

This thesis presented various machine learning models for detecting mental well-being states relevant in industrial HRC. The contribution in the development and deployment of these models can be clubbed into the following three steps:

- **Pre-processing**: For models based on images, the pre-processing steps were fairly straight forward and involved face-crop and scaling to the input dimensions of the neural network. Appropriate data augmentation methods (e.g., rotation, horizontal flip) were applied to improve variations in input data.

  For physiological signals, noise removal techniques (e.g., frequency filter) were implemented based on the literature. Additionally, relevant features were extracted for training shallow models. To mitigate the influence of person-specific characteristics of physiological responses, feature-wise MinMax normalization was implemented.

- **Training Models**: Image-based deep learning models (VGG16) were implemented for attention recognition, emotion recognition, and pain detection. ECG-based deep learning models (e.g., DeepECGNet) were trained for detecting stress. HRV-based shallow models (e.g., simple feed-forward neural network) were developed for predicting stress and flow states. The hyper parameters for each model were determined through empirical evaluations.

- **Real-time Detection**: Plugins were developed to incorporate the pre-processing steps and the trained models into the SSI framework. This enabled creating SSI pipelines which facilitated the prediction of these states in real-time.

9.2.3 Analysis-based Insights

Each chapter from Part 2 and 3 of this thesis presented insights based on the findings of the corresponding studies that contributed to the research in the field. Some of these insights were obtained often through quantitative analyses and are listed below:

- **XAI-based Learned Feature Representation Comparison**: Chapter 4 presented a XAI-based approach to compare the learned feature representations of two image-based deep learning models. This approach was demonstrated in two use cases. First, it was used to identify the features forgotten while transfer learning pain detection from emotion recognition model. Analysis showed that the forgotten features were facial regions that do not occur on prototypical pain expression patterns. Second, the approach was used to identify differences in expressions of clinical and experimental pain with the help of automatically learned feature representations. Although there were no distinguishing features, the relevance of features varied depending on the pain dataset.

- **Behavioral Patterns of Neurotypical and ASD Participants**: Chapter 6 presented the qualitative and quantitative analyses conducted on the long-term HRC dataset, which resulted in insights about the behavioral patterns exhibited by neurotypical and ASD participants. While some behavioral patterns (e.g., manifestations of tiredness/boredom) were similar in both groups, they differed in other aspects such as prioritizing cobot, duration of gaze towards cobot, and assembly routine. Some of the behavioral patterns in ASD participants were akin to social interactions, although industrial HRC cannot be considered an obvious social scenario.

- **Relationship between Flow States and Facial/Physiological Responses**: In Chapter 7, the flow in industry-like HRC dataset was analyzed to determine if facial emotion estimation and HRV features differed significantly during different flow-related states (boredom, anxiety, flow). Features like heart rate and mean HRV showed significant differences, with trends similar to existing studies on flow during mentally demanding tasks.
- **Gaze Cues for Initiating Joint Activities**: Chapter 8 analyzed the gaze behaviors of participants, especially around the joint activity phase of an industrial HRC task. The participants showed a tendency to look at the cobot when they were nearly ready for the joint activity, making it a potential cue for initiating joint activities.
- **Differences in Gaze Cues of Neurotypical and ASD Participants**: A key insight from Chapter 8 was regarding the differences between neurotypical and ASD participants interacting with a cobot with automatic gaze-based triggers. The ASD participant frequently triggered the cobot considerably early, before they were ready for the joint activity. This observation led to a hypothesis that ASD participants might not utilize gaze to initiate joint activity but for a different purpose (e.g., for monitoring the status of the cobot). However, further research is required for verifying this hypothesis.

9.2.4 Performance-based Insights

Some of the insights were obtained by evaluating model performances (e.g., accuracy, F1-score) and are presented below:

- **Source Datasets for Transfer Learning**: Chapter 3 demonstrated the potential of gaze estimation as a source model for transfer learning attention recognition. This was demonstrated through two separate datasets, where the models yielded high performances. Similarly, Chapter 4 demonstrated emotion recognition as a source model for transfer learning pain detection.
- **Need for Natural Distractions**: Chapter 3 presented the results of cross-dataset evaluation of models trained on a dataset with guided gaze when applied to images of participants performing industry-like HRC task. The slight drop in the recall of distraction class indicates the need for additional modalities. Moreover, it emphasizes the importance of evaluating attention recognition models using natural manifestations of distraction instead of posed distraction.
- **Lack of Generalizability of ECG-based Deep Learning Stress Models**: A notable observation in Chapter 5 was the lack of generalization capabilities of deep learning

stress models trained on ECG signals. Moreover, increasing training data by combining datasets resulted in a slight reduction in performance compared to individual dataset models. These observations point to a plausible reliance of these models on the recording hardware.

- **Factors Influencing Generalizability of Stress Models**: One of the findings in Chapter 5 centered around the factors that influence the generalization capabilities of a HRV-based stress models. Through multiple cross-dataset evaluations, the impact of factors such as stress eliciting technique, intensity, and sensor hardware were assessed. Stressor type was found to be an important factor in developing generalizable HRV models that can be applied to other contexts.

Chapter 10

Outlook

While this thesis investigated the development of models applicable to industrial Human-Robot Collaboration (HRC) settings, the states considered (attention/distraction, pain, stress, flow) are by no means an exhaustive list. Other states, such as fatigue and trust, have been identified in the literature as relevant in industrial HRC scenarios. For instance, Coronado et al. [2022] highlights the importance of detecting fatigue for worker safety, while Baltrusch et al. [2022] discusses the role of trust in HRC scenarios. These states are equally important for improving worker well-being, and similar challenges regarding model applicability and generalizability of features are likely to be encountered when exploring their detection using machine learning. Hence, there is a need to extend the investigations presented in this thesis to these states as well.

The realization of well-being-friendly cobots extends beyond model development. The findings of this thesis contribute to laying a foundation for worker state detection in industrial HRC settings. A few potential avenues for future work utilizing these models are discussed below.

10.1 Applying Models in Actual Factories

The industrial HRC studies presented in this thesis were conducted in a laboratory setting. The laboratory setup enabled the inclusion of ASD participants, but it does not address the applicability of findings to actual factories. Although the task and scenarios were designed to closely mimic an industrial workcell, certain aspects such as environmental and organizational factors could not be effectively simulated in a lab. For instance, actual factories may have higher noise levels as there are multiple robots or workcells operating simultaneously. The elevated noise levels may act as a stimuli (e.g., stress stimuli) that influences the workers' state. Moreover, a real-life factory worker's experience profile and skills are likely different from the participants from the studies. The participants were neither familiar with cobots nor had experience with manufacturing jobs.

Given the limitations of the lab setup, there is an evident need for investigating real-life worker experiences. As a first step, a focus group interview was conducted to obtain the perspectives of different stakeholders of a company. The interview involved three employees - a cobot worker (male), a learning&development manager (female), and an ergonomics&health manager (male) - from a car manufacturing company in Germany. The interview was conducted following the story interview method outlined by Mackay [2023]. This method involves asking the participant to provide a walk-through of a recent experience followed by specific questions to delve deeper into certain aspects of the experience.

The interview took place at the AI production network facility in Augsburg, Germany. While the story interview method typically focus on real-life events and experiences, this focus group interview investigated the effectiveness of challenge-skill balance (associated with flow) in an HRC task. This choice aimed to elicit a variety of states (boredom, stress, flow), rather than a specific single state (e.g., pain, distraction). Since participants' factory does not have the flow model that controls cobot behavior, a lab setup was utilized to enable the participants to experience varying cobot behavior. To this end, a modified version of the setup described in Chapter 7 was devised at the facility in Augsburg. The HRC scenario was scripted to include a slow (low cobot production speed), fast (high cobot production speed), and adaptive (cobot production speed adjusted to operator) conditions at specific intervals.

After the participants engaged in the scripted scenario, they participated in the focus group interview. The story interview explored their experience with the task, their impressions of the HRC system, and their perspectives on similar situations within their factory setting. The participants' responses (originally in German) were transcribed (using Whisper[1]) and translated (using DeepL[2]) for analysis.

The participants reported that the monotonous nature of the task mimicked real-world scenarios, which made the task feel realistic (ergonomics&health manager: "*[...] it was realistic because that's really how a production could be*"). The cobot worker, in particular, expressed a desire for a feature that allows the cobot to operate at different production rates – something currently unavailable in their existing cobot system. Moreover, all participants agreed that a slow cobot could lead to mind wandering and distraction (cobot worker: " *[...] when you're working on something like [this monotonous job] you're not focused*"). They suggested that the cobot system could monitor such states and recommend breaks or job rotations to maintain worker engagement. Unlike the participants from the study in Chapter 7, the fast condition, which resulted in cobot waiting for operator to finish assembly, did not lead to the cobot worker feeling pressured or anxious.

The insights from the focus group interview support the validity of the setup used in the studies presented in Part 3 of this thesis. Some of the findings may directly be applicable to actual factories, while others may need further investigation and refinement before deploying in real-world industrial HRC settings. Nevertheless, there is a need for further investigations conducted in actual factories involving cobot workers.

10.2 Cobot Adaptations

One promising avenue for utilizing the worker state detection models lies in automatic cobot adaptations. The adaptations would be triggered by the model predictions, allowing

[1] https://github.com/openai/whisper

[2] https://www.deepl.com/translator

the cobot to adjust its behavior in real-time to promote worker well-being. A few cobot adaptations were explored conceptually in Chapters 5, 7, and 8 through illustrative cartoon strips. Chapter 8 further demonstrated the feasibility of this approach by implementing a system that leveraged the attention recognition model to trigger the cobot for joint activity based on the operator's cues. Ideally, future cobots should be capable of responding to a broader range of worker cues for enhanced collaboration.

A critical step in developing adaptive cobot systems involves designing suitable cobot behaviors for each detected state. For instance, how should a cobot respond if it detects operator stress? The optimal response likely depends on the context and specific task characteristics. If the cobot's proximity causes stress, an appropriate action might involve adjusting the cobot's trajectory to maintain a more comfortable distance. However, this adaptation needs to consider the specific task requirements and work cell layout. Nevertheless, designing adaptive cobot behaviors for worker well-being requires further research through user studies to explore optimal responses for various detected states.

A practical consideration for cobot adaptations based on real-time worker state predictions is identifying the appropriate adaptation window. As noted by Nunnari et al. [2023], while models predict a state at a high frequency, cobot behavior adjustments should occur at a suitable rate. If adaptations happen too frequently, they can disrupt the smoothness of interaction. Conversely, infrequent adaptations might lead to missed opportunities to address the worker state. Finding the right balance between responsiveness and smoothness is crucial for effective cobot adaptation.

For promoting an inclusive work environment, the machine learning models need to be validated for a diverse population. For example, individuals with Autism Spectrum Disorder (ASD) may exhibit blunted stress responses to social evaluation compared to neurotypical individuals [Corbett et al., 2019]. This raises important questions: Can models trained on data from neurotypical individuals be applied to ASD workers? Is there a need for separate models tailored to detecting worker states in ASD populations? Furthermore, cobot adaptations themselves need to be designed with inclusivity in mind. As Chapter 8 demonstrated, adaptations designed for a neurotypical population might not be suitable for ASD workers. Further research is necessary to tailor cobot adaptations based on the individual needs of workers.

10.3 Cobot Interventions

In recent years, digital well-being interventions have gained significant traction across various settings [Armaou et al., 2020; Ferrari et al., 2022]. For example, Howe et al. [2022] designed a chatbot-based intervention system to mitigate stress in the workplace. In an industrial HRC setting, cobots equipped with worker-state detection models could potentially deliver targeted interventions to address negative states. Imagine a scenario where a cobot detects that a worker is experiencing boredom. The cobot could then offer to take over some repetitive tasks, potentially reducing boredom. Examples of such interventions were conceptually explored in Chapters 3 and 4 using illustrative scenarios.

Like cobot adaptations, designing suitable interventions targeted at specific negative states is crucial. Additionally, determining the timing and method of intervention delivery is equally important. Howe et al. investigated the effectiveness of scheduled versus adaptive intervention timings for stress reduction. While their study did not reveal a significant difference in stress reduction, participants preferred having some agency in scheduling interventions over fully automated approaches. Beyond timing, the communication medium also needs careful consideration. Industrial environments can be noisy, making audio-based communication challenging. Therefore, it is essential to explore alternative intervention modalities suited for industrial settings.

The research regarding model applicability to diverse populations is equally relevant for cobot interventions. Furthermore, just as with cobot adaptations, interventions should be designed with inclusivity in mind to accommodate the needs of individual workers. For example, interventions that rely on color-coded information (e.g., red lights signifying a stop command) might not be suitable for workers with color blindness. Future research should explore methods for personalizing cobot interventions to ensure effectiveness for a broad range of users.

10.4 Explainable AI

The widespread use of machine learning models in industrial HRC raises concerns about their interpretability. These models are often criticized for being “black boxes", where users struggle to understand how the model arrives at its decisions [Hassija et al., 2024]. In industrial HRC settings, explainable AI (XAI) techniques are valuable tools in not only developing robust models but also enhancing the collaboration experience.

One key benefit of XAI is its ability to help identify and address potential biases within the models. In Chapter 4, for instance, XAI techniques were used to identify biases in a pain detection model that stemmed from behavioral differences in the training datasets. However, bias can also arise from a lack of variation within the data. For example, dataset biases could lead to a pain detection model learning wrinkled faces (a common sign of aging) as an indicator of pain due to dataset bias. Such a model would be biased against older workers. By leveraging XAI techniques, model developers can select training datasets and parameters that promote fairness and mitigate bias.

Beyond bias detection, XAI plays a significant role in building trust in cobot technology, which could lead to improved user acceptance among workers. As highlighted by Wang et al. [2018], simple explanations involving the prediction confidence of the model are sufficient to increase the user’s trust in the prediction. Furthermore, XAI can help alleviate concerns about perceived unpredictability caused by changes in cobot behavior. For example, a cobot that adjusts its speed based on the operator’s proximity might be perceived as erratic or unpredictable. The operator might then feel the need to constantly monitor the cobot’s movements, increasing their cognitive load. However, the cobot explaining its reason for adaptation (e.g., “I am adjusting my speed because your stress levels appear elevated at higher speeds") can potentially reduce perceived unpredictability.

Furthermore, XAI can foster a greater willingness to accept interventions [Kuhl et al., 2020]. Imagine a scenario where the cobot suggests a change in its configuration due to a detected pain state in the operator during a joint activity. The operator might be more receptive to this suggestion if the cobot explains the reasoning behind its recommendation.

It is also necessary to consider how these explanations can be tailored for cobot workers, potentially using visualizations or simplified language to enhance understanding for users who may not have a technical background.

10.5 Virtual Characters

As discussed in Chapter 1, the introduction of cobots can lead to a reduction in human-human interaction within the work cell. One approach for mitigating this effect involves leveraging virtual characters to recreate some aspects of social interaction. Virtual characters have been successfully employed in various applications to represent social roles, such as well-being coaches [El Kamali et al., 2020], tutors [Armando et al., 2022], and training partners [Bosman et al., 2019]. For instance, Arora et al. [2022] proposes a socially aware virtual character that acts as a physical therapy assistant, motivating patients during at-home exercise routines.

In an industrial HRC scenario, a virtual cobot companion can serve as a social interface for the physical cobot, making it feel more like a teammate. Research by Nicora et al. [2023] supports this concept, demonstrating that participants attributed social presence to a virtual character introduced into a collaborative work cell, even assigning it social roles like colleague or supervisor.

The concept of social presence is often linked with the social facilitation effect [Park and Catrambone, 2007]. Social facilitation describes the phenomenon where individuals perform better on well-learned or easy tasks when in the presence of others. Conversely, social inhibition occurs when the presence of others hinders performance on complex or novel tasks. Research has shown that social facilitation can also be observed in the presence of virtual characters, not just humans [Sterna et al., 2019]. As evidenced by the results presented in Table 7.4, a typical HRC assembly task is perceived as less demanding and low-effort. Moreover, in a real industrial scenario, the workers often get well-versed in their tasks over time. So, introducing a virtual character has the potential to improve worker productivity through social facilitation.

The models developed in this thesis can be leveraged to control the virtual character's behavior, further enhancing social presence and potentially amplifying the social facilitation effect [von der Pütten et al., 2010]. For example, when the model detects the operator looking at the cobot, the virtual character could establish eye contact to acknowledge the operator's cue. Further research is necessary to explore how the virtual character's behavior should be adapted based on specific detected states to optimize social presence and enhance social facilitation.

Beyond social facilitation, virtual characters can also play a valuable role in communicating interventions and explaining the cobot's decisions related to adaptations and recommended actions. Imagine a scenario where the model detects boredom in the operator. The cobot might adapt its work pace to increase the challenge level of the task. However, an unexpected change in the cobot's behavior could be confusing or lead to the perception of an unpredictable cobot. In this situation, the virtual character could inform the operator that the cobot is adjusting its work pace to make the task more engaging.

10.6 Ethical Considerations

Sense, Plan, and Act are the fundamental elements of AI in an HRC scenario [Murphy, 2019; Neupane et al., 2024]. This thesis focuses on the sensing aspect, centering around collecting data from human participants and training machine learning models to improve HRC. Since human data is involved, it is necessary to address the ethical concerns and best practices for this area of research. The existing literature has proposed various dimensions of ethical considerations involving AI models [Greene et al., 2019; Lo Piano, 2020; Batliner et al., 2020; Ximenes and Ramalho, 2021]. This section leverages some of the commonly discussed principles to guide the discussion within the context of human data collection and well-being-friendly cobots.

- **Beneficence and Non-Maleficence**: These two interconnected principles are fundamental to ethical research. Beneficence emphasizes using data for benefitting the user (e.g., improving well-being), while non-maleficence refers to ensuring the data and resulting technologies do not cause harm. For instance, using worker state detection results (e.g., fatigue, stress) to make employment decisions (e.g., job termination, promotions) would violate both principles by reducing workers' mental well-being and harming their livelihood.

  In this thesis, the training datasets were collected from participants in a laboratory environment and not from actual workers. Moreover, the pipelines employing these models facilitate controlling the cobot's behavior in real-time, with no provision for storing the detection results.

- **Privacy and Data Protection**: Ensuring user privacy and protecting confidential information is paramount when working with human data. The typical ethical practices to address these concerns include data anonymization measures (e.g., pseudonymization) and compliance with relevant data protection regulations, such as the General Data Protection Regulation (GDPR). Cloud-based storage and computing are becoming increasingly popular with large amounts of data and intensive computations. This further necessitates anonymization and GDPR compliance as the data gets uploaded to an external server.

  To address the privacy concerns, the data collected in this thesis are shared by strictly following the permissions granted through informed consent. Additionally, to protect the identities of the ASD participants, their facial images are blurred (both in this thesis and associated publications).

- **Transparency and Explainability**: With the growing complexity of machine learning models, understanding the rationale behind a model's prediction becomes increasingly challenging. Simpler models like Random Forest Classifiers and Support Vector Machines rely on extracted features, making feature importance calculations relatively straightforward. Feature importance indicates which features have the most significant influence on the model's decision. However, in deep learning models, the features are learned during training and may not be readily available. To address this challenge, researchers have developed XAI techniques to generate visualizations that aid in interpreting the learned features and their contribution to specific model decisions.

Overall, incorporating explanations in industrial HRC settings has multiple benefits. Some of these benefits were discussed with examples in the previous section (see Section 10.4).

- **Bias and Fairness**: One significant advantage of employing XAI techniques is their ability to help identify potential biases within the models. Biased models can lead to discriminatory decision-making that disproportionately affects certain groups. These biases often stem from biases inherent in the training data. So, to mitigate biases in the models, it is crucial to utilize diverse datasets that represent a broader demographic.

  The datasets utilized in this thesis have limitations in terms of diversity. While the male-female gender ratio might be acceptable in most cases, the datasets lack sufficient representation of non-binary individuals. Additionally, participant recruitment primarily occurred through universities or academic institutions, limiting the age range and potentially introducing educational bias. Furthermore, the participant population is predominantly European nationals. However, it is worth highlighting that the inclusion of participants with ASD in the long-term study represents a positive step towards a more diverse dataset.

- **Human Autonomy**: In the context of AI models and HRC, human autonomy refers to respecting the worker's decision-making authority over the AI's suggestions, even if the model suggests a seemingly "optimal" choice. Worker autonomy should be prioritized, even if it hinders autonomous cobot behaviors. For instance, a cobot system might recommend a break when it detects worker fatigue. However, the worker's decision to continue working should be respected, even if taking a break could improve productivity.

  Another aspect of autonomy to consider is the ability of the user to opt out of engaging with the AI technology. A good ethical practice would be letting workers control what data they share with the system. This could involve options to opt out of specific data collection modalities (e.g., physiological data, facial expressions) or entirely from the cobot's AI-driven interventions. For instance, a worker who prefers self-directed fatigue management might choose not to receive break suggestions from the cobot.

  This thesis did not delve deeply into exploring methods for facilitating human autonomy. A basic version of opting out was practiced during data acquisition, where the participants could withdraw from the study at any point and request to delete their data.

This section discussed a few good ethical practices and how they can be incorporated into industrial HRC scenarios. For a comprehensive overview of best practices in developing AI models, refer to the works of Lo Piano [2020], Batliner et al. [2020], and Ximenes and Ramalho [2021].

Appendix A

Questionnaires

A.1 Flow Short Scale

1. I feel just the right amount of challenge
   ○ ○ ○ ○ ○ ○ ○ ○

2. My thoughts/activities run fluidly and smoothly
   ○ ○ ○ ○ ○ ○ ○ ○

3. I don't notice time passing
   ○ ○ ○ ○ ○ ○ ○ ○

4. I have no difficulty concentrating
   ○ ○ ○ ○ ○ ○ ○ ○

5. My mind is completely clear
   ○ ○ ○ ○ ○ ○ ○ ○

6. I am totally absorbed in what I am doing
   ○ ○ ○ ○ ○ ○ ○ ○

7. The right thoughts/movements occur of their own accord
   ○ ○ ○ ○ ○ ○ ○ ○

8. I know what I have to do each step of the way
   ○ ○ ○ ○ ○ ○ ○ ○

9. I feel that I have everything under control
   ○ ○ ○ ○ ○ ○ ○ ○

10. I am completely lost in thought
    ○ ○ ○ ○ ○ ○ ○ ○

A.2 NASA-TLX Task Load Questionnaire

Mental Demand **How much mental activity was required?**

0 20

Very low Very high

Physical Demand **How much physical activity was required?**

0 20

Very low Very high

Temporal Demand **How much time pressure did you feel due to the task's pace?**

0 20

Very low Very high

Performance **How successful were you in accomplishing the task?**

0 20

Very low Very high

Effort **How hard did you work to achieve your performance?**

0 20

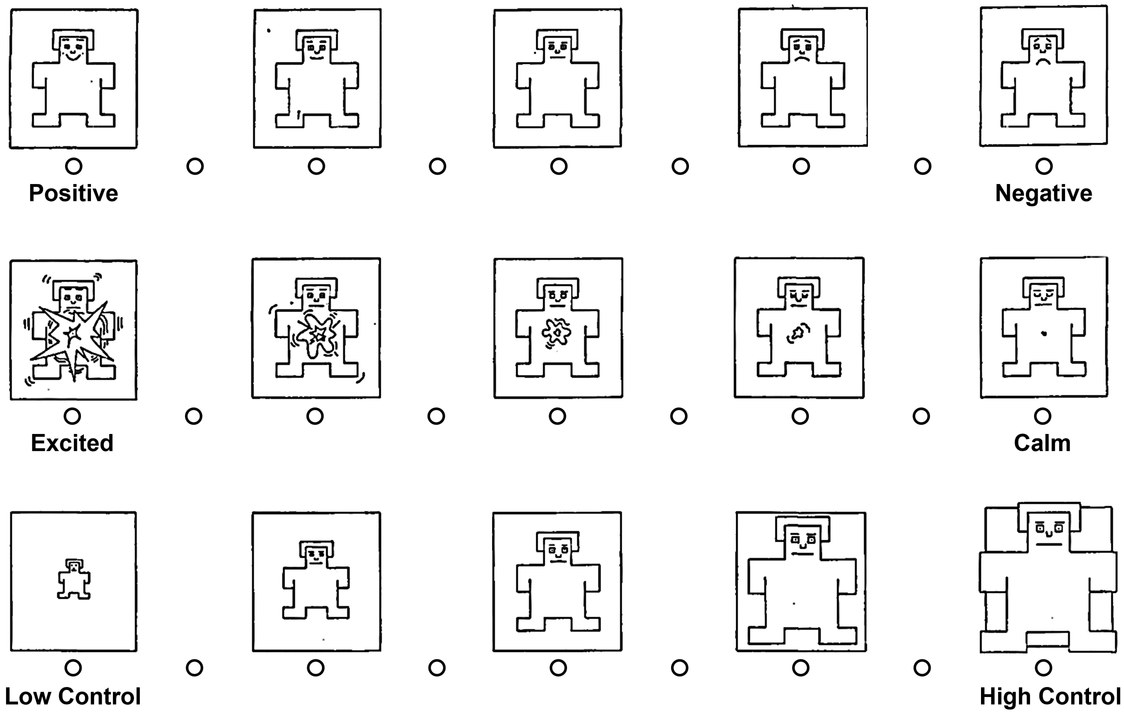
Very low Very high

Frustration **How insecure, discouraged, irritated, stressed, and annoyed were you?**

0 20

Very low Very high

A.3 SAM Emotion Questionnaire



Note: The pictorial representations are adapted from Bradley and Lang [1994] with permission from the publisher.

A.4 ELoC Locus of Control Questionnaire

I worked hard to achieve the performance I wanted

1 5

Completely Disagree — Completely Agree

If I did a good performance, it was because of me

1 5

Completely Disagree — Completely Agree

I did the task because I felt like doing it and not because I was asked to do it

1 5

Completely Disagree — Completely Agree

Appendix B

Usage of AI and Third-Party Content

- Some elements (e.g., top-view of the cobot) of Figures 3.7, 6.4, 7.3, and 8.5 were generated using Hugging Face[1] text-to-image models.

- Some elements (e.g., cobot icon) of Figures 2.5, 3.1, 3.7, 4.1, 4.7, 5.1, 5.5, 6.4, 7.1, 7.3, 8.1, and 8.5 use icons from Flaticon[2], which are available for free.

- The contents of this thesis were written without any use of generative AI. However, the original sentences were slightly rephrased using online large language models - ChatGPT[3] and Gemini[4]. I manually verified the rephrased content before incorporating them into the thesis.

- I used Grammarly[5] to check for grammatical errors.

[1] https://huggingface.co/

[2] https://www.flaticon.com/

[3] https://chatgpt.com/

[4] https://gemini.google.com/app

[5] https://app.grammarly.com/

Appendix C

Academic Activities

C.1 Publications

1. P. Prajod, D. Schiller, T. Huber, and E. André. Do deep neural networks forget facial action units?—Exploring the effects of transfer learning in health related facial expression recognition. *AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action*, 1013:217, 2022b

2. P. Prajod, T. Huber, and E. André. Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions. In *International Conference on Multimedia Modeling*, pages 311–322. Springer, 2022a

3. R. Arora, M. L. Nicora, P. Prajod, D. Panzeri, E. André, P. Gebhard, and M. Malosio. Employing socially interactive agents for robotic neurorehabilitation training. *arXiv preprint arXiv:2206.01587*, 2022

4. P. Prajod and E. André. On the generalizability of ECG-based stress detection models. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 549–554. IEEE, 2022

5. A. Heimerl, P. Prajod, S. Mertes, T. Baur, M. Kraus, A. Liu, H. Risack, N. Rohleder, E. André, and L. Becker. ForDigitStress: A multi-modal stress dataset employing a digital job interview scenario. *arXiv preprint arXiv:2303.07742*, 2023

6. P. Prajod, M. Lavit Nicora, M. Malosio, and E. André. Gaze-based attention recognition for human-robot collaboration. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 140–147, 2023a

7. M. Mondellini, P. Prajod, M. L. Nicora, M. Chiappini, E. Micheletti, F. A. Storm, R. Vertechy, E. André, and M. Malosio. Behavioral patterns in robotic collaborative assembly: Comparing neurotypical and autism spectrum disorder participants. *Frontiers in Psychology*, 14, 2023

8. F. Nunnari, M. L. Nicora, P. Prajod, S. Beyrodt, L. Chehayeb, E. André, P. Gebhard, M. Malosio, and D. Tsovaltzi. Understanding and mapping pleasure, arousal and

dominance social signals to robot-avatar behavior. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2023

9. S. Beyrodt, M. L. Nicora, F. Nunnari, L. Chehayeb, P. Prajod, T. Schneeberger, E. André, M. Malosio, P. Gebhard, and D. Tsovaltzi. Socially interactive agents as cobot avatars: Developing a model to support flow experiences and well-being in the workplace. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2023

10. M. Mondellini, M. L. Nicora, P. Prajod, E. André, R. Vertechy, A. Antonietti, and M. Malosio. Exploring the dynamics between cobot's production rhythm, locus of control and emotional state in a collaborative assembly scenario. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2024

11. P. Prajod, M. L. Nicora, M. Mondellini, G. Tauro, R. Vertechy, M. Malosio, and E. André. Gaze detection and analysis for initiating joint activity in industrial human-robot collaboration. *arXiv preprint arXiv:2312.06643*, 2023b

12. P. Prajod, M. Lavit Nicora, M. Mondellini, M. M. Falerni, R. Vertechy, M. Malosio, and E. André. Flow in human-robot collaboration—Multimodal analysis and perceived challenge detection in industrial scenarios. *Frontiers in Robotics and AI*, 11:1393795, 2024a

13. P. Prajod, B. Mahesh, and E. André. Stressor type matters!–Exploring factors influencing cross-dataset generalizability of physiological stress detection. *arXiv preprint arXiv:2405.09563*, 2024b

14. R. Arora, P. Prajod, M. L. Nicora, D. Panzeri, G. Tauro, R. Vertechy, M. Malosio, E. André, and P. Gebhard. Socially interactive agents for robotic neurorehabilitation training: Conceptualization and proof-of-concept study. *arXiv preprint arXiv:2406.12035*, 2024

15. M. Lavit Nicora, P. Prajod, M. Mondellini, G. Tauro, R. Vertechy, E. André, and M. Malosio. Gaze detection as a social cue to initiate natural human-robot collaboration in an assembly task. *Frontiers in Robotics and AI*, 11:1394379, 2024

16. P. Prajod, D. Schiller, D. W. Don, and E. André. Faces of experimental pain: Transferability of deep learned heat pain features to electrical pain. *arXiv preprint arXiv:2406.11808*, 2024c

C.2 Awards

- Winner of AI4Pain Grand Challenge at 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) 2024

C.3 Reviews

- ACM International Conference on Multimodal Interaction (ICMI 2024) – 2 papers
- International Conference on Affective Computing and Intelligent Interaction (ACII 2024) – 2 papers
- International Journal of Social Robotics, ISSN: 1875-4791 (2023) – 2 articles
- IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2020) – 1 paper

C.4 Talks

- *Studying User Experiences in HRC* at MindBot Public Event, CNR-Lecco, Italy (September 2023)
- *Using AI to Promote Good Mental Health and Well-being in Industry 5.0* at KI-Produktionsnetzwerk, Augsburg, Germany (June 2023)
- *Integrating People Characterized by ASD to Industry 5.0* at Seminar in University of Augsburg, Germany (November 2022)

C.5 Press Coverage

- *Cobots of the Future – Promoting Mental Health in the Workplace* in Research in Bavaria Website (February 2024) [1]
- *Kollege Roboter: Uni Augsburg untersucht Zusammenarbeit von Menschen und Cobots (Robot Colleague: University of Augsburg Investigates Collaboration between Humans and Cobots)* in Aichacher Zeitung (January 2024) [2]
- *Forschung an der Universität: Künstliche Intelligenz soll Schmerzen erkennen (Research at the University: Artificial Intelligence to Detect Pain)* in Augsburger Allgemeine (April 2022) [3]
- *Mit künstlicher Intelligenz zum gesünderen Arbeitsplatz? (Using Artificial Intelligence to Create Healthier Workplace?)* in University Presse (February 2022) [4]

[1] https://www.research-in-bavaria.de/de/future-of-work/cobots-and-mental-health

[2] https://www.aichacher-zeitung.de/kollege-roboter-uni-augsburg-untersucht-zusammenarbeit-von-m
cnt-id-ps-6a0b87d3-3aa7-45d0-9e3d-2c1a6030386d

[3] https://www.augsburger-allgemeine.de/augsburg/augsburg-forschung-an-der-universitaet-kuenstli
html

[4] https://www.uni-augsburg.de/de/campusleben/neuigkeiten/2022/02/25/5850/

Bibliography

W. M. S. Abedi, A. T. Sadiq, and I. Nadher. Modified CNN-LSTM for pain facial expressions recognition. *International Journal of Advanced Science and Technology*, 29(3 Special Issue):304 – 312, 2020.

Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa. Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498*, 2017.

B. Adak and S. Halder. Systematic review on prevalence for autism spectrum disorder with respect to gender and socio-economic status. *Journal of Mental Disorders and Treatment*, 3(1):1–9, 2017.

R. Adattil, P. Thorvald, and D. Romero. Assessing the psychosocial impacts of Industry 4.0 technologies adoption in the Operator 4.0: Literature review & theoretical framework. *International Journal of Industrial Engineering and Management*, 15(1):59–80, 2024.

A. Adel. Future of Industry 5.0 in society: Human-centric solutions, challenges and prospective research areas. *Journal of Cloud Computing*, 11(1):40, 2022.

H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

S. S. Agati, R. D. Bauer, M. da S. Hounsell, and A. S. Paterno. Augmented reality for manual assembly in Industry 4.0: Gathering guidelines. In *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*, pages 179–188. IEEE, 2020.

C. Ahlstrom, K. Kircher, and A. Kircher. A gaze-based driver distraction warning system and its effect on visual behavior. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):965–973, 2013.

Z. Ahmad, S. Rabbani, M. R. Zafar, S. Ishaque, S. Krishnan, and N. Khan. Multi-level stress assessment from ECG in a virtual reality environment using multimodal fusion. *IEEE Sensors Journal*, 2023.

A. W. Aidoo. The influence of stress on the health of workers in manufacturing industry. *Inquiry-Sarajevo Journal of Social Science*, 2(2):65–75, 2016.

J. Aigrain, S. Dubuisson, M. Detyniecki, and M. Chetouani. Person-specific behavioural features for automatic stress detection. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 3, pages 1–6. IEEE, 2015.

A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib. Progress and prospects of the human–robot collaboration. *Autonomous Robots*, 42:957–975, 2018.

V. Akçit and E. Barutçu. The relationship between performance and loneliness at workplace: A study on academicians. *European Scientific Journal, Special Issue*, pages 235–243, 2017.

H. Akechi, A. Senju, H. Uibo, Y. Kikuchi, T. Hasegawa, and J. K. Hietanen. Attention to eye contact in the west and east: Autonomic responses and evaluative ratings. *PloS one*, 8 (3):e59312, 2013.

T. Åkerstedt, A. Knutsson, P. Westerholm, T. Theorell, L. Alfredsson, and G. Kecklund. Mental fatigue, work and sleep. *Journal of psychosomatic Research*, 57(5):427–433, 2004.

N. K. Al-Qazzaz, I. F. Abdulazez, and S. A. Ridha. Simulation recording of an ECG, PCG, and PPG for feature extractions. *Al-Khwarizmi Engineering Journal*, 10(4):81–91, 2014.

M. Albaladejo-González, J. A. Ruipérez-Valiente, and F. Gómez Mármol. Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):11011–11021, 2023.

M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans. iNNvestigate neural networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC, 2013.

D. Andronas, G. Apostolopoulos, N. Fourtakas, and S. Makris. Multi-modal interfaces for natural human-robot interaction. *Procedia Manufacturing*, 54:197–202, 2021.

T. Arai, R. Kato, and M. Fujita. Assessment of operator stress induced by robot collaboration in assembly. *CIRP annals*, 59(1):5–8, 2010.

X. Arakaki, R. J. Arechavala, E. H. Choy, J. Bautista, B. Bliss, C. Molloy, D.-A. Wu, S. Shimojo, Y. Jiang, M. T. Kleinman, et al. The connection between heart rate variability (HRV), neurological health, and cognition: A literature review. *Frontiers in Neuroscience*, 17:1055445, 2023.

J. Arents, V. Abolins, J. Judvaitis, O. Vismanis, A. Oraby, and K. Ozols. Human–robot collaboration trends and safety aspects: A systematic review. *Journal of Sensor and Actuator Networks*, 10(3):48, 2021.

M. Armando, M. Ochs, and I. Régner. The impact of pedagogical agents' gender on academic learning: A systematic review. *Frontiers in Artificial Intelligence*, 5:862997, 2022.

M. Armaou, S. Konstantinidis, and H. Blake. The effectiveness of digital interventions for psychological well-being in the workplace: A systematic review protocol. *International journal of environmental research and public health*, 17(1):255, 2020.

R. Arora, M. L. Nicora, P. Prajod, D. Panzeri, E. André, P. Gebhard, and M. Malosio. Employing socially interactive agents for robotic neurorehabilitation training. *arXiv preprint arXiv:2206.01587*, 2022.

R. Arora, P. Prajod, M. L. Nicora, D. Panzeri, G. Tauro, R. Vertechy, M. Malosio, E. André, and P. Gebhard. Socially interactive agents for robotic neurorehabilitation training: Conceptualization and proof-of-concept study. *arXiv preprint arXiv:2406.12035*, 2024.

R. Arya, J. Singh, and A. Kumar. A survey of multidisciplinary domains contributing to affective computing. *Computer Science Review*, 40:100399, 2021.

S. M. Asish, E. Hossain, K. AK, and C. W. Borst. Deep learning on eye gaze data to classify student distraction level in an educational VR environment. In *International Conference on Artificial Reality and Telexistence Eurographics Symposium on Virtual Environments (ICAT-EGVE)*, 2021.

G. Atherton and L. Cross. Seeing more than human: Autism and anthropomorphic theory of mind. *Frontiers in Psychology*, 9:528, 2018.

M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, et al. The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal EmoPain dataset. *IEEE Transactions on Affective Computing*, 7(4):435–451, 2015.

S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, and J. D. Zhang. An introduction to machine learning. *Clinical pharmacology & therapeutics*, 107(4):871–885, 2020.

K. Baek, S. Yang, M. Lee, and I. Chung. The association of workplace psychosocial factors and musculoskeletal pain among korean emotional laborers. *Safety and health at work*, 9(2):216–223, 2018.

R. Bailey and M. Clarke. Meanings and models of stress and coping. In *Stress and Coping in Nursing*, pages 3–34. Springer, 1989.

A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, S. Sturmbauer, E.-M. Meßner, B. M. Kudielka, et al. An evaluation of speech-based recognition of emotional and physiological markers of stress. *Frontiers in Computer Science*, 3:750284, 2021.

A. Bajaj. Performance metrics in machine learning, 2023. URL `https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide`. Accessed: 2024-06-03.

S. Baltrusch, F. Krause, A. de Vries, W. van Dijk, and M. de Looze. What about the human in human robot collaboration? A literature review on HRC's effects on aspects of job quality. *Ergonomics*, 65(5):719–740, 2022.

K. Baraka, M. Couto, F. S. Melo, A. Paiva, and M. Veloso. "Sequencing Matters": Investigating suitable action sequences in robot-assisted autism therapy. *Frontiers in Robotics and AI*, 9, 2022.

Y. Baştanlar and M. Özuysal. Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, pages 105–128, 2014.

A. Batliner, S. Hantke, and B. Schuller. Ethics and good practice in computational paralinguistics. *IEEE Transactions on Affective Computing*, 13(3):1236–1253, 2020.

D. Battini, N. Berti, S. Finco, I. Zennaro, and A. Das. Towards Industry 5.0: A multi-objective job rotation model for an inclusive workforce. *International Journal of Production Economics*, 250:108619, 2022.

A. Bauer, D. Wollherr, and M. Buss. Human–robot collaboration: A survey. *International Journal of Humanoid Robotics*, 5(01):47–66, 2008.

T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André. NovA: Automated analysis of nonverbal signals in social interactions. In *Human Behavior Understanding: 4th International Workshop, HBU 2013, Barcelona, Spain, October 22, 2013. Proceedings 4*, pages 160–171. Springer, 2013.

V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann. Blazeface: Sub-millisecond neural face detection on mobile GPUs. *arXiv preprint arXiv:1907.05047*, 2019.

L. Becker, A. Heimerl, and E. André. ForDigitStress: Presentation and evaluation of a new laboratory stressor using a digital job interview-scenario. *Frontiers in Psychology*, 14: 1182959, 2023.

B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad. A transformer architecture for stress detection from ECG. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 132–134, 2021.

A. Belardinelli. Gaze-based intention estimation: Principles, methodologies, and applications in HRI. *ACM Transactions on Human-Robot Interaction*, 2023.

M. Benchekroun, D. Istrate, V. Zalc, and D. Lenne. A multi-modal dataset (MMSD) for acute stress bio-markers. In *International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 377–392. Springer, 2022.

M. Benchekroun, P. E. Velmovitsky, D. Istrate, V. Zalc, P. P. Morita, and D. Lenne. Cross dataset analysis for generalizability of HRV-based stress detection models. *Sensors*, 23 (4):1807, 2023.

S. Beyrodt, M. L. Nicora, F. Nunnari, L. Chehayeb, P. Prajod, T. Schneeberger, E. André, M. Malosio, P. Gebhard, and D. Tsovaltzi. Socially interactive agents as cobot avatars: Developing a model to support flow experiences and well-being in the workplace. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2023.

Y. Bian, C. Yang, F. Gao, H. Li, S. Zhou, H. Li, X. Sun, and X. Meng. A framework for physiological indicators of flow in VR games: Construction and preliminary evaluation. *Personal and Ubiquitous Computing*, 20:821–832, 2016.

T. Billah, S. M. Rahman, M. O. Ahmad, and M. Swamy. Recognizing distractions for assistive driving by tracking body parts. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(4):1048–1062, 2018.

J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani. A non-EEG biosignals dataset for assessment and visualization of neurological status. In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, pages 110–114. IEEE, 2016.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, 2006. ISBN 9781493938438.

G. Blandino, F. Montagna, and M. Cantamessa. Workload and stress evaluation in advanced manufacturing systems. *Materials Research Proceedings*, 35, 2023.

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

K. Bosman, T. Bosse, and D. Formolo. Virtual agents for professional social skills training: An overview of the state-of-the-art. In *Intelligent Technologies for Interactive Entertainment: 10th EAI International Conference, INTETAIN 2018, Guimarães, Portugal, November 21-23, 2018, Proceedings 10*, pages 75–84. Springer, 2019.

P. J. Bota, C. Wang, A. L. Fred, and H. P. da Silva. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access*, 7:140990–141020, 2019.

J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey, and J. Ventre-Dominey. I reach faster when I see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics*, 6:3, 2012.

D. Bouhassira, N. Danziger, N. Atta, and F. Guirimand. Comparison of the pain suppressive effects of clinical and experimental painful conditioning stimuli. *Brain*, 126(5):1068–1078, 2003.

M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1): 49–59, 1994.

S. Bragança, E. Costa, I. Castellucci, and P. M. Arezes. A brief overview of the use of collaborative robots in Industry 4.0: Human role and safety. *Occupational and environmental safety and health*, pages 641–650, 2019.

S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack. Machine recognition and representation of neonatal facial displays of acute pain. *Artificial intelligence in medicine*, 36(3): 211–222, 2006.

S. Brahnam, L. Nanni, S. McMurtrey, A. Lumini, R. Brattin, M. Slack, and T. Barrier. Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from gaussian of local descriptors. *Applied Computing and Informatics*, 19(1/2): 122–143, 2023.

J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, 49(1):1017–1034, 2013.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

C. Brod. *Technostress : The human cost of the computer revolution.* Addison-Wesley, 1984.

A. Burns and J. Tulip. Detecting flow in games using facial expressions. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 45–52. IEEE, 2017.

M. Buss, D. Carton, B. Gonsior, K. Kuehnlenz, C. Landsiedel, N. Mitsou, R. de Nijs, J. Zlotowski, S. Sosnowski, E. Strasser, et al. Towards proactive human-robot interaction in human environments. In *2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 1–6. IEEE, 2011.

J. Bütepage and D. Kragic. Human-robot collaboration: From psychology to social robotics. *arXiv preprint arXiv:1705.10146*, 2017.

J. A. Caldwell, J. L. Caldwell, L. A. Thompson, and H. R. Lieberman. Fatigue and its management in the workplace. *Neuroscience & Biobehavioral Reviews*, 96:272–289, 2019.

A. L. Callara, L. Sebastiani, N. Vanello, E. P. Scilingo, and A. Greco. Parasympathetic-sympathetic causal interactions assessed by time-varying multivariate autoregressive modeling of electrodermal activity and heart-rate-variability. *IEEE Transactions on Biomedical Engineering*, 68(10):3019–3028, 2021.

L. Camaioni. Mind knowledge in infancy: The emergence of intentional communication. *Early Development and Parenting*, 1(1):15–22, 1992.

S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma. PPG and EDA dataset collected with Empatica E4 for stress assessment. *Data in Brief*, 53:110102, 2024.

T. S. Campbell, J. A. Johnson, and K. A. Zernicke. Gate control theory of pain. In *Encyclopedia of behavioral medicine*, pages 914–916. Springer, 2020.

A. Campeau-Lecours, U. Côté-Allard, D.-S. Vu, F. Routhier, B. Gosselin, and C. Gosselin. Intuitive adaptive orientation control for enhanced human–robot interaction. *IEEE Transactions on robotics*, 35(2):509–520, 2018.

J. Campisi, Y. Bravo, J. Cole, and K. Gobeil. Acute psychosocial stress differentially influences salivary endocrine and immune measures in undergraduate students. *Physiology & Behavior*, 107(3):317–321, 2012. ISSN 0031-9384.

Y. S. Can, B. Arnrich, and C. Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, 92:103139, 2019.

Y. S. Can, B. Mahesh, and E. André. Approaches, applications, and challenges in physiological emotion recognition—A tutorial overview. *Proceedings of the IEEE*, 111(10): 1287–1313, 2023.

F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

R. Cañigueral and A. F. de C. Hamilton. The role of eye gaze during natural social interactions in typical and autistic people. *Frontiers in Psychology*, 10:560, 2019.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

N. Caporusso, T. Cao, and B. Thaman. Comparative analysis of RGB-based eye-tracking for large-scale human-machine applications. *Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems*, 22(22), 2022.

L. P. Carlini, L. A. Ferreira, G. A. Coutrin, V. V. Varoto, T. M. Heiderich, R. C. Balda, M. C. Barros, R. Guinsburg, and C. E. Thomaz. A convolutional neural network-based mobile application to bedside neonatal pain assessment. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 394–401. IEEE, 2021.

L. P. Carlini, G. de A. S. Coutrin, L. A. Ferreira, J. do C. A. Soares, G. V. T. Silva, T. M. Heiderich, R. de C. X. Balda, M. C. de Moraes Barros, R. Guinsburg, and C. E. Thomaz. Human vs machine towards neonatal pain assessment: A comprehensive analysis of the facial features extracted by health professionals, parents, and convolutional neural networks. *Artificial Intelligence in Medicine*, 147:102724, 2024.

J. Carr, B. Kelley, R. Keaton, and C. Albrecht. Getting to grips with stress in the workplace: Strategies for promoting a healthier, more productive environment. *Human Resource Management International Digest*, 19(4):32–38, 2011.

G. Cartella, M. Cornia, V. Cuculo, A. D'Amelio, D. Zanca, G. Boccignone, and R. Cucchiara. Trends, applications, and challenges in human attention modelling. *arXiv preprint arXiv:2402.18673*, 2024.

M. S. Cary. The role of gaze in the initiation of conversation. *Social Psychology*, 41(3): 269–271, 1978.

R. T. Chadalavada, H. Andreasson, M. Schindler, R. Palm, and A. J. Lilienthal. Accessing your navigation plans! Human-Robot intention transfer using eye-tracking glasses. In *Advances in Manufacturing Technology XXXII*, pages 253–258. IOS Press, 2018.

W. P. Chan, M. Crouch, K. Hoang, C. Chen, N. Robinson, and E. Croft. Design and implementation of a human-robot joint action framework using augmented reality and eye gaze. *arXiv preprint arXiv:2208.11856*, 2022.

E. Charlton. Ethical guidelines for pain research in humans. Committee on ethical issues of the International Association for the Study of Pain. *Pain*, 63(3):277–278, 1995.

J. Charron, P. Rainville, and S. Marchand. Direct comparison of placebo effects on clinical and experimental pain. *The Clinical journal of pain*, 22(2):204–211, 2006.

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

J. Chen, Z. Chi, and H. Fu. A new approach for pain event detection in video. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 250–254. IEEE, 2015.

J. Chen, Z. Chi, and H. Fu. A new framework with multiple tasks for detecting and locating pain events in video. *Computer Vision and Image Understanding*, 155:113–123, 2017.

W. Chen, S. Zheng, and X. Sun. Introducing MDPSD, a multimodal dataset for psychological stress detection. In *Big Data: 8th CCF Conference, BigData 2020, Chongqing, China, October 22–24, 2020, Revised Selected Papers 8*, pages 59–82. Springer, 2021.

Y. Cheng, H. Wang, Y. Bao, and F. Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Chesterfield PCT Service. *The gate control theory of pain.* NHS Foundation Trust, 2014.

H.-M. Cho, H. Park, S.-Y. Dong, and I. Youn. Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network. *Sensors*, 19(20):4408, 2019.

F. Christandl, K. Mierke, and C. Peifer. Time flows: Manipulations of subjective time progression affect recalled flow and performance in a subsequent task. *Journal of Experimental Social Psychology*, 74:246–256, 2018.

R. Ciuffini, V. Cofini, M. Muselli, S. Necozione, A. Piroli, and A. Marrelli. Emotional arousal and valence in patients with fibromyalgia: A pilot study. *Frontiers in Pain Research*, 4: 1075722, 2023.

E. A. Clark, J. Kessinger, S. E. Duncan, M. A. Bell, J. Lahne, D. L. Gallagher, and S. F. O'Keefe. The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review. *Frontiers in Psychology*, 11:507534, 2020.

L. Claudia, I. Oscar, P.-G. Héctor, and V. J. Marco. Poincaré plot indexes of heart rate variability capture dynamic adaptations after haemodialysis in chronic renal failure patients. *Clinical physiology and functional imaging*, 23(2):72–80, 2003.

M. Cohen, J. Quintner, and S. van Rysewyk. Reconsidering the international association for the study of pain definition of pain. *Pain reports*, 3(2):e634, 2018.

J. F. Cohn and F. De la Torre. Automated face analysis for affective computing. *The Oxford handbook of affective computing*, page 131, 2014.

B. E. Cole. Pain management: Classifying, understanding, and treating pain. *Hospital physician*, pages 23–30, 2002.

T. Colombino, D. Gallo, S. Shreepriya, Y. Im, and S. Cha. Ethical design of a robot platform for disabled employees: Some practical methodological considerations. *Frontiers in Robotics and AI*, 8:643160, 2021.

B. A. Corbett, R. A. Muscatello, and C. Baldinger. Comparing stress and arousal systems in response to different social contexts in children with ASD. *Biological psychology*, 140: 119–130, 2019.

E. Coronado, T. Kiyokawa, G. A. G. Ricardez, I. G. Ramirez-Alpizar, G. Venture, and N. Yamanobe. Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an Industry 5.0. *Journal of Manufacturing Systems*, 63:392–410, 2022.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

B. Coşkun, S. Ay, D. Erol Barkana, H. Bostanci, İ. Uzun, A. B. Oktay, B. Tuncel, and D. Tarakci. A physiological signal database of children with different special needs for stress recognition. *Scientific data*, 10(1):382, 2023.

G. A. Coutrin, L. P. Carlini, L. A. Ferreira, T. M. Heiderich, R. C. Balda, M. C. Barros, R. Guinsburg, and C. E. Thomaz. Convolutional neural networks for newborn pain assessment using face images: A quantitative and qualitative comparison. In *International Conference on Medical Imaging and Computer-Aided Diagnosis*, pages 503–513. Springer, 2022.

M. R. Cowie, J. I. Blomster, L. H. Curtis, S. Duclaux, I. Ford, F. Fritz, S. Goldman, S. Janmohamed, J. Kreuzer, M. Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9, 2017.

M. Csikszentmihalyi. *Beyond boredom and anxiety.* San Francisco: Jossey-Bass, 1975.

M. Csikszentmihalyi. *Finding flow: The psychology of engagement with everyday life.* Hachette UK, 2020.

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

L. Cunha, D. Silva, and S. Maggioli. Exploring the status of the human operator in Industry 4.0: A systematic review. *Frontiers in Psychology*, 13:889129, 2022.

J. Czakon. F1 score vs ROC AUC vs accuracy vs PR AUC: Which evaluation metric should you choose?, 2023. URL `https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc`. Accessed: 2024-06-03.

L. Dai, J. Broekens, and K. P. Truong. Real-time pain detection in facial expressions for health robotics. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 277–283. IEEE, 2019.

Z. Dai, J. Park, A. Kaszowska, and C. Li. Detecting worker attention lapses in human-robot interaction: An eye tracking and multimodal sensing study. In *2023 28th International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE, 2023.

I. Damian, M. Dietz, and E. André. The SSJ framework: Augmenting social interactions using mobile signal processing and live feedback. *Frontiers in ICT*, 5:13, 2018.

O. Damm, K. Malchus, P. Jaecks, S. Krach, F. Paulus, M. Naber, A. Jansen, I. Kamp-Becker, W. Einhaeuser-Treyer, P. Stenneken, et al. Different gaze behavior in human-robot interaction in Asperger's syndrome: An eye-tracking study. In *2013 IEEE RO-MAN*, pages 368–369. IEEE, 2013.

S. D'Angelo and B. Schneider. Shared gaze visualizations in collaborative interactions: Past, present and future. *Interacting with Computers*, 33(2):115–133, 2021.

K. Das, M. Papakostas, K. Riani, A. Gasiorowski, M. Abouelenien, M. Burzo, and R. Mihalcea. Detection and recognition of driver distraction using multimodal signals. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–28, 2022.

A.-M. D'Cruz, M. E. Ragozzino, M. W. Mosconi, S. Shrestha, E. H. Cook, and J. A. Sweeney. Reduced behavioral flexibility in autism spectrum disorders. *Neuropsychology*, 27(2):152, 2013.

M. F. de Sampaio Barros, F. M. Araújo-Moreira, L. C. Trevelin, and R. Radel. Flow experience and the mobilization of attentional resources. *Cognitive, Affective, & Behavioral Neuroscience*, 18:810–823, 2018.

V. De Simone, V. Di Pasquale, V. Giubileo, and S. Miranda. Human-robot collaboration: An analysis of worker's performance. *Procedia Computer Science*, 200:1540–1549, 2022.

Z. Degutyte and A. Astell. The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, 12:616471, 2021.

K. Delgado, J. M. Origgi, T. Hasanpoor, H. Yu, D. Allessio, I. Arroyo, W. Lee, M. Betke, B. Woolf, and S. A. Bargal. Student engagement dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3628–3636, 2021.

Ö. de Manzano, T. Theorell, L. Harmat, and F. Ullén. The psychophysiology of flow during piano playing. *Emotion*, 10(3):301, 2010.

P. J. Dewe, M. P. O'Driscoll, and C. L. Cooper. Theories of psychological stress at work. *Handbook of occupational health and wellness*, pages 23–38, 2012.

E. Di Lascio, S. Gashi, M. E. Debus, and S. Santini. Automatic recognition of flow during work activities using context and physiological signals. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

V. Di Pasquale, V. De Simone, V. Giubileo, and S. Miranda. A taxonomy of factors influencing worker's performance in human–robot collaboration. *IET Collaborative Intelligent Manufacturing*, 5(1):e12069, 2023.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

H. K. Dishar and L. A. Muhammed. A review of the overfitting problem in convolution neural network and remedy approaches. *Journal of Al-Qadisiyah for computer science and mathematics*, 15(2):155–165, 2023.

W. O. dos Santos, D. Dermeval, L. B. Marques, I. I. Bittencourt, S. Isotani, and I. F. Silveira. Flow theory to promote learning in educational systems: Is it really relevant? *Revista Brasileira de Informática na Educação*, 26(2), 2018.

F. Doshi-Velez and B. Kim. Considerations for evaluation and generalization in interpretable machine learning. *Explainable and interpretable models in computer vision and machine learning*, pages 3–17, 2018.

M. Dubey and L. Singh. Automatic emotion recognition using facial expression: A review. *International Research Journal of Engineering and Technology (IRJET)*, 3(2):488–492, 2016.

K. Dufour, J. Ocampo-Jimenez, and W. Suleiman. Visual–spatial attention as a comfort measure in human–robot collaborative tasks. *Robotics and Autonomous Systems*, 133: 103626, 2020.

P. C. Duttweiler. The internal control index: A newly developed measure of locus of control. *Educational and psychological measurement*, 44(2):209–221, 1984.

J. Dwyer, R. Gomez, J. Donovan, and C. Brophy. *Advancing Human-Robot Interaction Within Industrial Settings; What role might social cues play in the future of robotic interface design?* BMW Group+ QUT Design Academy, 2021.

J. L. Edens and K. M. Gil. Experimental induction of pain: Utility in the study of clinical pain. *Behavior Therapy*, 26(2):197–216, 1995.

S. G. Edwards, L. J. Stephenson, M. Dalmaso, and A. P. Bayliss. Social orienting in gaze leading: A mechanism for shared attention. *Proceedings of the Royal Society B: Biological Sciences*, 282(1812):20151141, 2015.

J. Egede, M. Valstar, M. T. Torres, and D. Sharkey. Automatic neonatal pain estimation: An acute pain in neonates database. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

P. Ekman and K. G. Heider. The universality of a contempt expression: A replication. *Motivation and emotion*, 12(3):303–308, 1988.

M. El Kamali, L. Angelini, M. Caon, F. Carrino, C. Röcke, S. Guye, G. Rizzo, A. Mastropietro, M. Sykora, S. Elayan, et al. Virtual coaches for older adults' wellbeing: A systematic review. *IEEE Access*, 8:101884–101902, 2020.

I. El Makrini, S. A. Elprama, J. Van den Bergh, B. Vanderborght, A.-J. Knevels, C. I. Jewell, F. Stals, G. De Coppel, I. Ravyse, J. Potargent, et al. Working with Walt: How a cobot was developed and inserted on an auto assembly line. *IEEE Robotics & Automation Magazine*, 25(2):51–58, 2018.

S. El Zaatari, M. Marei, W. Li, and Z. Usman. Cobot programming for collaborative industrial tasks: An overview. *Robotics and Autonomous Systems*, 116:162–180, 2019.

O. Eldardeer, J. Gonzalez-Billandon, L. Grasse, M. Tata, and F. Rea. A biological inspired cognitive framework for memory-based multi-sensory joint attention in human-robot interactive tasks. *Frontiers in Neurorobotics*, 15:648595, 2021.

M. Elgendi, M. Jonkman, and F. De Boer. Frequency bands effects on QRS detection. *Biosignals*, 2003:2002, 2010.

M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PloS one*, 8(10):e76585, 2013.

E. Emerson, N. Fortune, G. Llewellyn, and R. Stancliffe. Loneliness, social support, social isolation and wellbeing among working age adults with and without disability: Cross-sectional study. *Disability and health journal*, 14(1):100965, 2021.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2015.

M. Faccio, I. Granata, A. Menini, M. Milanese, C. Rossato, M. Bottin, R. Minto, P. Pluchino, L. Gamberini, G. Boschetti, et al. Human factors in cobot era: A review of modern production systems features. *Journal of Intelligent Manufacturing*, 34(1):85–106, 2023.

T. Faibish, A. Kshirsagar, G. Hoffman, and Y. Edan. Human preferences for robot eye gaze in human-to-robot handovers. *International Journal of Social Robotics*, 14(4):995–1012, 2022.

J. Fan, P. Zheng, and S. Li. Vision-based holistic scene understanding towards proactive human–robot collaboration. *Robotics and Computer-Integrated Manufacturing*, 75: 102304, 2022.

X. Fan and S. Straube. Reporting on work-related low back pain: Data sources, discrepancies and the art of discovering truths. *Pain management*, 6(6):553–559, 2016.

R. Fang, R. Zhang, E. Hosseini, A. M. Parenteau, S. Hang, S. Rafatirad, C. E. Hostinar, M. Orooji, and H. Homayoun. Prevent over-fitting and redundancy in physiological signal analyses for stress detection. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2585–2588. IEEE, 2022.

A. Fedorova, Z. Dvorakova, and H. Atas. Workplace well-being in employee estimates. In *Proceedings of the International Scientific Conference "Smart Nations: Global Trends In The Digital Economy" Volume 1*, pages 529–540. Springer, 2022.

R. Fernandes-Magalhaes, A. Carpio, D. Ferrera, D. Van Ryckeghem, I. Peláez, P. Barjola, M. E. De Lahoz, M. C. Martín-Buro, J. A. Hinojosa, S. Van Damme, et al. Pain E-motion faces database (PEMF): Pain-related micro-clips for emotion research. *Behavior Research Methods*, 55(7):3831–3844, 2023.

M. Ferrari, S. Allan, C. Arnold, D. Eleftheriadis, M. Alvarez-Jimenez, A. Gumley, and J. F. Gleeson. Digital interventions for psychological well-being in university students: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 24(9):e39686, 2022.

L. A. Ferreira, L. P. Carlini, G. de Almeida Sá Coutrin, T. M. Heideirich, M. C. de Moraes Barros, R. Guinsburg, and C. E. Thomaz. Revisiting N-CNN for clinical practice. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 231–240. Springer, 2023.

F. Ferri, G. C. Campione, R. Dalla Volta, C. Gianelli, and M. Gentilucci. Social requests and social affordances: How they affect the kinematics of motor sequences during interactions between conspecifics. *PloS one*, 6(1):e15855, 2011.

K. Fischer, L. C. Jensen, F. Kirstein, S. Stabinger, Ö. Erkent, D. Shukla, and J. Piater. The effects of social gaze in human-robot collaborative assembly. In *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*, pages 204–213. Springer, 2015.

C. Florea, L. Florea, and C. Vertan. Learning pain from emotion: Transferred HoT data representation for pain intensity estimation. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*, pages 778–790. Springer, 2015.

A. Frischen, A. P. Bayliss, and S. P. Tipper. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694, 2007.

C. Fullagar, A. Delle Fave, and S. Van Krevelen. Flow at work: The evolution of a construct. In *Current Issues in Work and Organizational Psychology*, pages 278–299. Routledge, 2018.

J. Gabriel, O. Peretemode, and D. Dinges. Industrial fatigue: A workman's great enemy. *Iosr Journal Of Business And Management (Iosr-Jbm)*, 20(10):9–14, 2018.

A. Gacek. An introduction to ECG signal processing and analysis. In *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence*, pages 21–46. Springer, 2011.

A. Gaggioli, P. Cipresso, S. Serino, and G. Riva. Psychophysiological correlates of flow during daily activities. *Annual Review of Cybertherapy and Telemedicine*, 191:65–69, 2013.

G. M. Galvani, S. Korivand, A. Ajoudani, J. Gong, and N. Jalili. A framework for human-robot teaming performance prediction: Reinforcement learning and eye movement analysis. In *ASME International Mechanical Engineering Congress and Exposition*, volume 87608, page V003T03A068. American Society of Mechanical Engineers, 2023.

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

P. Gebhard, G. Mehlmann, and M. Kipp. Visual scenemaker—A tool for authoring interactive virtual characters. *Journal on Multimodal User Interfaces*, 6(1-2):3–11, 2012.

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

S. Gedam and S. Paul. Automatic stress detection using wearable sensors and machine learning: A review. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2020.

R. Gervasi, K. Aliev, L. Mastrogiacomo, and F. Franceschini. User experience and physiological response in human-robot collaboration: A preliminary investigation. *Journal of Intelligent & Robotic Systems*, 106(2):36, 2022.

R. Gervasi, F. Barravecchia, L. Mastrogiacomo, and F. Franceschini. Applications of affective computing in human-robot interaction: State-of-art and challenges for manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 237(6-7):815–832, 2023.

D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva. Using activity-related behavioural features towards more effective automatic stress detection. *PLoS ONE*, 7(9):1–16, 2012. doi: 10.1371/journal.pone.0043571.

A. Giallanza, G. La Scalia, R. Micale, and C. M. La Fata. Occupational health and safety issues in human-robot collaboration: State of the art and open challenges. *Safety science*, 169:106313, 2024.

G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 2019.

G. Giannakakis, M. R. Koujan, A. Roussos, and K. Marias. Automatic stress detection evaluating models of facial action units. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*, pages 728–733. IEEE, 2020.

G. Giannakakis, M. R. Koujan, A. Roussos, and K. Marias. Automatic stress analysis from facial videos based on deep facial action units recognition. *Pattern Analysis and Applications*, pages 1–15, 2022.

C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter. Vocal indices of stress: A review. *Journal of voice*, 27(3):390–e21, 2013.

S. W. Gilroy, M. Cavazza, and M. Benayoun. Using affective trajectories to describe states of flow in interactive art. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology*, pages 165–172, 2009.

T. Giraud, B. Ravenet, C. Tai Dang, J. Nadel, E. Prigent, G. Poli, E. André, and J.-C. Martin. “Can you help me move this over there?”: Training children with ASD to joint action through tangible interaction and virtual agent. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 1–12, 2021.

S. Gkikas and M. Tsiknakis. Automatic assessment of pain based on deep learning methods: A systematic review. *Computer methods and programs in biomedicine*, 231:107365, 2023a.

S. Gkikas and M. Tsiknakis. A full transformer-based framework for automatic pain estimation using videos. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–6. IEEE, 2023b.

B. Gladysz, T.-A. Tran, D. Romero, T. van Erp, J. Abonyi, and T. Ruppert. Current development on the Operator 4.0 and transition towards the Operator 5.0: A systematic literature review in light of Industry 5.0. *Journal of Manufacturing Systems*, 70:160–185, 2023.

S. Glasauer, M. Huber, P. Basili, A. Knoll, and T. Brandt. Interacting in time and space: Investigating human-human and human-robot joint action. In *19th international symposium in robot and human interactive communication*, pages 252–257. IEEE, 2010.

B. Gleeson, K. MacLean, A. Haddadi, E. Croft, and J. Alcazar. Gestures for industry intuitive human-robot communication from human observation. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 349–356. IEEE, 2013.

S. Goldman, C. Wang, M. W. Salgado, P. E. Greene, M. Kim, and I. Rapin. Motor stereotypies in children with autism and other developmental disorders. *Developmental Medicine & Child Neurology*, 51(1):30–38, 2009.

C. Gomez Cubero and M. Rehm. Intention recognition in human robot interaction based on eye tracking. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*, pages 428–437. Springer, 2021.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

J. Goris, M. Brass, C. Cambier, J. Delplanque, J. R. Wiersema, and S. Braem. The relation between preference for predictability and autistic traits. *Autism Research*, 13(7):1144–1154, 2020.

O. C. Görür, B. Rosman, F. Sivrikaya, and S. Albayrak. Social cobots: Anticipatory decision-making for collaborative robots incorporating unexpected human behaviors. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 398–406, 2018.

S. Gradl. *The Stroop Room: A Wearable Virtual Reality Stress Laboratory Based on the Electrocardiogram*. FAU University Press, 2020. Doctoral Thesis.

A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2015.

A. Greco, G. Valenza, J. Lázaro, J. M. Garzón-Rey, J. Aguiló, C. de la Cámara, R. Bailón, and E. P. Scilingo. Acute stress state classification based on electrodermal activity modeling. *IEEE Transactions on Affective Computing*, 14(1):788–799, 2021.

S. A. Green, M. Billinghurst, X. Chen, and J. G. Chase. Human-robot collaboration: A literature review and augmented reality approach in design. *International journal of advanced robotic systems*, 5(1):1, 2008.

D. Greene, A. L. Hoffmann, and L. Stark. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2122–2131. Hawaii International Conference on System Sciences, 2019.

E. H. Grosse, F. Sgarbossa, C. Berlin, and W. P. Neumann. Human-centric production and logistics system design and management: Transitioning from Industry 4.0 to Industry 5.0. *International Journal of Production Research*, 61(22):7749–7759, 2023.

E. Grossi, E. Caminada, M. Goffredo, B. Vescovo, T. Castrignano, D. Piscitelli, G. Valagussa, M. Franceschini, and F. Vanzulli. Patterns of restricted and repetitive behaviors in autism spectrum disorders: A cross-sectional video recording study. Preliminary report. *Brain sciences*, 11(6):678, 2021.

S. Gruss, M. Geiger, P. Werner, O. Wilhelm, H. C. Traue, A. Al-Hamadi, and S. Walter. Multi-modal signals for analyzing pain responses to thermal and electrical stimuli. *JoVE (Journal of Visualized Experiments)*, (146):e59057, 2019.

J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

S. Gu, F. Wang, N. P. Patel, J. A. Bourgeois, and J. H. Huang. A model for basic emotions using observations of behavior in drosophila. *Frontiers in Psychology*, 10:445286, 2019.

L. Gualtieri, E. Rauch, and R. Vidoni. Emerging research fields in safety and ergonomics in industrial collaborative robotics: A systematic literature review. *Robotics and Computer-Integrated Manufacturing*, 67:101998, 2021.

Y. Guo, G. Zhao, and M. Pietikäinen. Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Transactions on Image Processing*, 25(5): 1977–1992, 2016.

R. Gupta, A. Bhongade, and T. K. Gandhi. Multimodal wearable sensors-based stress and affective states prediction model. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 30–35. IEEE, 2023.

O. Guy-Evans. Peripheral nervous system (PNS): Parts and function, 2023. URL `https://www.simplypsychology.org/peripheral-nervous-system.html`. Accessed: 2024-05-28.

T. Hadjistavropoulos, K. D. Craig, S. Duck, A. Cano, L. Goubert, P. L. Jackson, J. S. Mogil, P. Rainville, M. J. Sullivan, A. C. de C. Williams, et al. A biopsychosocial formulation of pain communication. *Psychological bulletin*, 137(6):910, 2011.

T. Hadjistavropoulos, M. Browne, K. Prkachin, B. Taati, A. Ashraf, and A. Mihailidis. Pain in severe dementia: A comparison of a fine-grained assessment approach to an observational checklist designed for clinical settings. *European Journal of Pain*, 22(5):915–925, 2018.

A. F. de C. Hamilton. Gazing at me: The importance of social meaning in understanding direct-gaze cues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371 (1686):20150080, 2016.

H. Han, K. Byun, and H.-G. Kang. A deep learning-based stress detection algorithm with speech signal. In *proceedings of the 2018 workshop on audio-visual scene understanding for immersive multimedia*, pages 11–15, 2018.

M. A. Haque, R. B. Bautista, F. Noroozi, K. Kulkarni, C. B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, et al. Deep multimodal pain recognition: A database and comparison of spatio-temporal visual modalities. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 250–257. IEEE, 2018.

Y. Haque, R. S. Zawad, C. S. A. Rony, H. Al Banna, T. Ghosh, M. S. Kaiser, and M. Mahmud. State-of-the-art of stress prediction from heart rate variability using artificial intelligence. *Cognitive Computation*, 16(2):455–481, 2024.

K. Harman and P. Ruyak. Working through the pain: A controlled study of the impact of persistent pain on performing a computer task. *The Clinical journal of pain*, 21(3): 216–222, 2005.

L. Harmat, Ö. de Manzano, T. Theorell, L. Högman, H. Fischer, and F. Ullén. Physiological correlates of the flow experience during computer game playing. *International Journal of Psychophysiology*, 97(1):1–7, 2015.

S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

S. K. Hasnain, G. Mostafaoui, R. Salesse, L. Marin, and P. Gaussier. Intuitive human robot interaction based on unintentional synchrony: A psycho-experimental study. In *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7. IEEE, 2013.

T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid. Automatic detection of pain from facial expressions: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1815–1831, 2019.

V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.

K. Hayashi and I. Mizuuchi. Investigation of joint action: Eye blinking behavior improving human-robot collaboration. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1133–1139. IEEE, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

T. M. Heiderich, A. T. F. S. Leslie, and R. Guinsburg. Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. *Acta Paediatrica*, 104(2):e63–e69, 2015.

A. Heimerl, P. Prajod, S. Mertes, T. Baur, M. Kraus, A. Liu, H. Risack, N. Rohleder, E. André, and L. Becker. ForDigitStress: A multi-modal stress dataset employing a digital job interview scenario. *arXiv preprint arXiv:2303.07742*, 2023.

D. Hendricks. Employment and adults with autism spectrum disorders: Challenges and strategies for success. *Journal of vocational rehabilitation*, 32(2):125, 2010.

T. H. Holmes and R. H. Rahe. The social readjustment rating scale. *Journal of psychosomatic research*, 1967.

W. E. Hoogendoorn, M. N. van Poppel, P. M. Bongers, B. W. Koes, and L. M. Bouter. Systematic review of psychosocial factors at work and private life as risk factors for back pain. *Spine*, 25(16):2114–2125, 2000.

S. Hopko, J. Wang, and R. Mehta. Human factors considerations and metrics in shared space human-robot collaboration: A systematic review. *Frontiers in Robotics and AI*, 9: 799522, 2022.

A. Horvers, N. Tombeng, T. Bosse, A. W. Lazonder, and I. Molenaar. Detecting emotions through electrodermal activity in learning contexts: A systematic review. *Sensors*, 21 (23):7869, 2021.

D. Hovens. Workplace learning through human-machine interaction in a transient multilingual blue-collar work environment. *Journal of Linguistic Anthropology*, 30(3):369–388, 2020.

E. Howe, J. Suh, M. Bin Morshed, D. McDuff, K. Rowan, J. Hernandez, M. I. Abdin, G. Ramos, T. Tran, and M. P. Czerwinski. Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

C.-M. Huang and B. Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 83–90. IEEE, 2016.

B. Hwang, J. You, T. Vaessen, I. Myin-Germeys, C. Park, and B.-T. Zhang. Deep ECGNet: An optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. *TELEMEDICINE and e-HEALTH*, 24(10):753–772, 2018.

iMotions. Facial action coding system (FACS) - A visual guidebook, 2022. URL `https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/`. Accessed: 2024-05-24.

A. Innocenti, E. De Stefani, N. F. Bernardi, G. C. Campione, and M. Gentilucci. Gaze direction and request gesture in social interactions. *PLoS one*, 7(5):e36390, 2012.

T. Iqbal, P. Redon-Lurbe, A. J. Simpkin, A. Elahi, S. Ganly, W. Wijns, and A. Shahzad. A sensitivity analysis of biophysiological responses of stress for wearable sensors in connected health. *IEEE Access*, 9:93567–93579, 2021.

T. Iqbal, A. J. Simpkin, D. Roshan, N. Glynn, J. Killilea, J. Walsh, G. Molloy, S. Ganly, H. Ryman, E. Coen, et al. Stress monitoring using wearable sensors: A pilot study and stress-predict dataset. *Sensors*, 22(21):8135, 2022.

M. T. Irshad, F. Li, M. A. Nisar, X. Huang, M. Buss, L. Kloep, C. Peifer, B. Kozusznik, A. Pollak, A. Pyszka, et al. Wearable-based human flow experience recognition enhanced by transfer learning methods using emotion data. *Computers in Biology and Medicine*, 166: 107489, 2023.

S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti. Towards engagement models that consider individual factors in HRI: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task: Experiments with the iCub humanoid. *International Journal of Social Robotics*, 9:63–86, 2017.

S. A. Jackson and M. Csikszentmihalyi. *Flow in sports.* Human Kinetics, 1999.

S. A. Jackson, P. R. Thomas, H. W. Marsh, and C. J. Smethurst. Relationships between flow, self-concept, psychological skills, and performance. *Journal of applied sport psychology*, 13(2):129–153, 2001.

R. Jahanmahin, S. Masoud, J. Rickli, and A. Djuric. Human-robot interactions in manufacturing: A survey of human behavior modeling. *Robotics and Computer-Integrated Manufacturing*, 78:102404, 2022.

M. Jaiswal and C.-P. Bara. Muse: A multimodal dataset of stressed emotion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.

J. Jang, H. Shin, H. Aum, M. Kim, and J. Kim. Application of experiential locus of control to understand users' judgments toward useful experience. *Computers in Human Behavior*, 54:326–340, 2016.

C. Janiesch, P. Zschech, and K. Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD. *Signal Processing: Image Communication*, 88:115960, 2020.

S. Jha, N. Stogios, A. S. de Oliveira, S. Thomas, and R. P. Nolan. Getting into the zone: A pilot study of autonomic-cardiac modulation and flow state during piano performance. *Frontiers in Psychiatry*, 13:853733, 2022.

T. Jiang, J. L. Gradus, and A. J. Rosellini. Supervised machine learning: A brief primer. *Behavior therapy*, 51(5):675–687, 2020.

B. Johnston and P. de Chazal. A review of image-based automatic facial landmark identification techniques. *EURASIP Journal on Image and Video Processing*, 2018(1):86, 2018.

H. Jonsson and D. Persson. Towards an experiential model of occupational balance: An alternative perspective on flow theory analysis. *Journal of Occupational Science*, 13(1): 62–73, 2006.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

W. Ju. *The design of implicit interactions.* Morgan & Claypool Publishers, 2015.

M. Kächele, M. Amirian, P. Thiam, P. Werner, S. Walter, G. Palm, and F. Schwenker. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems*, 8:71–83, 2017.

K. Kaewkaisorn, K. Pintong, S. Bunyang, T. Tansawat, and T. Siriborvornratanakul. Student attentiveness analysis in virtual classroom using distraction, drowsiness and emotion detection. *Discover Education*, 3(1):1–14, 2024.

H. Kagermann and Y. Nonaka. *Revitalizing Human-Machine Interaction for the Advancement of Society: Perspectives from Germany and Japan.* Acatech-National Academy of Science and Engineering, 2019.

A. Kalatzis, S. Rahman, V. Girishan Prabhu, L. Stanley, and M. Wittie. A multimodal approach to investigate the role of cognitive workload and user interfaces in human-robot collaboration. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 5–14, 2023.

J. Kappesser and A. C. de Williams. Pain and negative emotions in the face: Judgements by health care professionals. *Pain*, 99(1):197–206, 2002.

A. Karan and A. Kaygun. Time series classification via topological data analysis. *Expert Systems with Applications*, 183:115326, 2021.

S. Karmakar, P. Singh, T. Varghese, M. B. Sheshachala, R. D. Gavas, R. K. Ramakrishnan, and A. Pal. CamTratak: RGB camera-based frugal real-time gaze tracker for real-world applications. *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 70–75, 2024.

R. Kato, M. Fujita, and T. Arai. Development of advanced cellular manufacturing system with human-robot collaboration. In *19th international symposium in robot and human interactive communication*, pages 355–360. IEEE, 2010.

J. Keller, H. Bless, F. Blomann, and D. Kleinböhl. Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology*, 47(4):849–852, 2011.

R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan. Measuring catastrophic forgetting in neural networks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26: 22–63, 1967.

K. H. Kendrick and J. Holler. Gaze direction signals response preference in conversation. *Research on Language and Social Interaction*, 50(1):12–32, 2017.

A. Khalid, P. Kirisci, Z. Ghrairi, K. Thoben, and J. Pannek. Towards implementing safety and security concepts for human-robot collaboration in the context of Industry 4.0. In *39th International MATADOR Conference on Advanced Manufacturing*, volume 2, pages 55–63, 2017.

M. S. Khalil and A. Elfaki. Stress theories. `https://www.slideshare.net/UDDent/theories-of-stress`, 2014. Accessed: 2024-05-21.

M. Khan, A. Haleem, and M. Javaid. Changes and improvements in Industry 5.0: A strategic approach to overcome the challenges of Industry 4.0. *Green Technologies and Sustainability*, 1(2):100020, 2023.

R. Kharghanian, A. Peiravi, and F. Moradi. Pain detection from facial images using unsupervised feature learning approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 419–422. IEEE, 2016.

R. Kher. Signal processing techniques for removing noise from ECG signals. *Journal of Biomedical Engineering and Research*, 3(101):1–9, 2019.

P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE international conference on computer vision workshops*, pages 19–27, 2015.

S. Khoshnoud, F. A. Igarzábal, and M. Wittmann. Peripheral-physiological and neural correlates of the flow experience while playing video games: A comprehensive review. *PeerJ*, 8:e10520, 2020.

J. Kildal, I. Maurtua, M. Martin, and I. Ipiña. Towards including workers with cognitive disabilities in the factory of the future. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 426–428, 2018.

B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018a.

H. Kim, J. K. Neubert, J. S. Rowan, J. S. Brahim, M. J. Iadarola, and R. A. Dionne. Comparison of experimental and acute clinical pain responses in humans as pain phenotypes. *The Journal of Pain*, 5(7):377–384, 2004.

H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry investigation*, 15(3):235, 2018b.

C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer. The 'Trier Social Stress Test'–A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

J. M. Kivikangas. *Psychophysiology of flow experience: An explorative study.* Helsingfors Universitet, 2006. Masters Thesis.

S. Klein, J. Huch, N. Reißner, P. Zwolsky, K. Weitz, M. Kraus, and E. André. Creating a framework for a user-friendly cobot failure management in human-robot collaboration. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 618–622, 2024.

C. M. Klingner and O. Guntinas-Lichius. Mimik und emotion. *Laryngo-rhino-otologie*, 102 (S 01):S115–S125, 2023.

M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal communication in human interaction*, volume 1. Holt, Rinehart and Winston New York, 1978.

M. T. Knierim, R. Rissler, V. Dorner, A. Maedche, and C. Weinhardt. The psychophysiology of flow: A systematic review of peripheral nervous system features. *Information Systems and Neuroscience: Gmunden Retreat on NeuroIS 2017*, pages 109–120, 2018.

M. T. Knierim, R. Rissler, A. Hariharan, M. Nadj, and C. Weinhardt. Exploring flow psychophysiology in knowledge work. In *Information Systems and Neuroscience: NeuroIS Retreat 2018*, pages 239–249. Springer, 2019.

B. C. Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2):401, 2018.

A. Kolbeinsson, E. Lagerstedt, and J. Lindblom. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. *Production & Manufacturing Research*, 7(1):448–471, 2019.

S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij. The SWELL knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*, pages 291–298, 2014.

S. Koldijk, M. A. Neerincx, and W. Kraaij. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on Affective Computing*, 9(2):227–239, 2016.

T. Kopp, M. Baumgartner, and S. Kinkel. Success factors for introducing industrial human-robot interaction in practice: An empirically driven framework. *The International Journal of Advanced Manufacturing Technology*, 112:685–704, 2021.

M. Koppenborg, P. Nickel, B. Naber, A. Lungfiel, and M. Huelke. Effects of movement speed and predictability in human–robot collaboration. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 27(4):197–209, 2017.

J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, et al. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1022–1040, 2019.

J. B. Kostis, A. Moreyra, M. Amendo, J. Di Pietro, N. Cosgrove, and P. Kuo. The effect of age on heart rate in subjects free of heart disease. Studies by ambulatory electrocardiography and maximal exercise stress test. *Circulation*, 65(1):141–145, 1982.

I. Kotseruba and J. K. Tsotsos. Attention for vision-based assistive and automated driving: A review of algorithms and datasets. *IEEE transactions on intelligent transportation systems*, 23(11):19907–19928, 2022.

K. J. Kovács, I. H. Miklós, and B. Bali. Psychological and physiological stressors. In *Techniques in the behavioral and neural sciences*, volume 15, pages 775–792. Elsevier, 2005.

N. Kuhl, J. Lobana, and C. Meske. Do you comply with AI?–Personalized explanations of learning algorithms and their impact on employees' compliance behavior. *arXiv preprint arXiv:2002.08777*, 2020.

B. Kühnlenz, M. Erhart, M. Kainert, Z.-Q. Wang, J. Wilm, and K. Kühnlenz. Impact of trajectory profiles on user stress in close human-robot interaction. *at-Automatisierungstechnik*, 66(6):483–491, 2018.

K. Kühnlenz, M. Westermann, and B. Kühnlenz. Impact of human gaze behavior and robot appearance on motion uncertainty during cooperative hand movement tasks. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1301–1306. IEEE, 2020.

D. Kulić and E. Croft. Physiological and subjective responses to articulated robot motion. *Robotica*, 25(1):13–27, 2007.

S. Kumawat, M. Verma, and S. Raman. LBVCNN: Local binary volume convolutional neural network for facial expression recognition from image sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

H. Kumazaki, T. Muramatsu, Y. Yoshikawa, Y. Matsumoto, H. Ishiguro, M. Kikuchi, T. Sumiyoshi, and M. Mimura. Optimal robot for intervention for individuals with autism spectrum disorders. *Psychiatry and Clinical Neurosciences*, 74(11):581–586, 2020.

M. Kunz and S. Lautenbacher. The faces of pain: A cluster analysis of individual differences in facial activity patterns of pain. *European Journal of Pain*, 18(6):813–823, 2014.

M. Kunz, J. Peter, S. Huster, and S. Lautenbacher. Pain and disgust: The facial signaling of two aversive bodily experiences. *PloS one*, 8(12):e83277, 2013.

M. Kunz, D. Seuss, T. Hassan, J. U. Garbas, M. Siebers, U. Schmid, M. Schöberl, and S. Lautenbacher. Problems of video-based pain detection in patients with dementia: A road map to an interdisciplinary solution. *BMC geriatrics*, 17(1):1–8, 2017.

M. Kunz, D. Meixner, and S. Lautenbacher. Facial muscle movements encoding pain—A systematic review. *Pain*, 160(3):535–549, 2019.

H. Kurniawan, A. V. Maslov, and M. Pechenizkiy. Stress detection from speech and galvanic skin response signals. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, pages 209–214. IEEE, 2013.

U. Kurylo and J. R. Wilson. Using human eye gaze patterns as indicators of need for assistance from a socially assistive robot. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*, pages 200–210. Springer, 2019.

Y.-M. Kwon, K.-W. Jeon, J. Ki, Q. M. Shahab, S. Jo, and S.-K. Kim. 3D gaze estimation and interaction to stereo display. *International Journal of Virtual Reality*, 5(3):41–45, 2006.

B. N. Kyle and D. W. McNeil. Autonomic arousal and experimentally induced pain: A critical review of the literature. *Pain Research and Management*, 19:159–167, 2014.

J. Lambert, J. Chapman, and D. Lurie. Challenges to the four-channel model of flow: Primary assumption of flow support the moderate challenging control channel. *The Journal of Positive Psychology*, 8(5):395–403, 2013.

M. Lavit Nicora, P. Prajod, M. Mondellini, G. Tauro, R. Vertechy, E. André, and M. Malosio. Gaze detection as a social cue to initiate natural human-robot collaboration in an assembly task. *Frontiers in Robotics and AI*, 11:1394379, 2024.

R. S. Lazarus. *Psychological stress and the coping process.* McGraw-Hill, 1966.

R. S. Lazarus and S. Folkman. *Stress, appraisal, and coping.* Springer publishing company, 1984.

R. S. Lazarus and S. Folkman. Transactional theory and research on emotions and coping. *European Journal of personality*, 1(3):141–169, 1987.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

M. Lee. Detecting affective flow states of knowledge workers using physiological sensors. *arXiv preprint arXiv:2006.10635*, 2020.

W. Lee, C. H. Park, S. Jang, and H.-K. Cho. Design of effective robotic gaze-based social cueing for users in task-oriented situations: How to overcome in-attentional blindness? *Applied Sciences*, 10(16):5413, 2020.

I. Lefter, G. J. Burghouts, and L. J. Rothkrantz. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Transactions on Affective Computing*, 7(2):162–175, 2015.

N. Leigh-Hunt, D. Bagguley, K. Bash, V. Turner, S. Turnbull, N. Valtorta, and W. Caan. An overview of systematic reviews on the public health consequences of social isolation and loneliness. *Public health*, 152:157–171, 2017.

C. Li, J. Zhai, and A. Barreto. Signal processing quantification of changes in the blood volume pulse (BVP) waveform due to exercise. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, volume 4, pages 3180–3183. IEEE, 2003.

S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

W. Li, Y. Hu, Y. Zhou, and D. T. Pham. Safe human–robot collaboration for industrial settings: A survey. *Journal of Intelligent Manufacturing*, pages 1–27, 2023.

L. Liakopoulos, N. Stagakis, E. I. Zacharaki, and K. Moustakas. CNN-based stress and emotion recognition in ambulatory settings. In *2021 12th international conference on information, intelligence, systems & applications (IISA)*, pages 1–8. IEEE, 2021.

S. Liao, L. Lin, and Q. Chen. Research on the acceptance of collaborative robots for the Industry 5.0 era – The mediating effect of perceived competence and the moderating effect of robot use self-efficacy. *International Journal of Industrial Ergonomics*, 95:103455, 2023.

A. Liapis, E. Faliagka, C. Katsanos, C. Antonopoulos, and N. Voros. Detection of subtle stress episodes during UX evaluation: Assessing the performance of the WESAD biosignals dataset. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*, pages 238–247. Springer, 2021.

H. Limaye and V. Deshmukh. ECG noise sources and various noise removal techniques: A survey. *International Journal of Application or Innovation in Engineering & Management*, 5(2):86–92, 2016.

M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2014.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

W. Lin and C. Li. Review of studies on emotion recognition and judgment based on physiological signals. *Applied Sciences*, 13(4):2573, 2023.

A. Lisowska, S. Wilk, and M. Peleg. Catching patient's attention at the right time to help them undergo behavioural change: Stress classification experiment from blood volume pulse. In *International Conference on Artificial Intelligence in Medicine*, pages 72–82. Springer, 2021.

Y. Liu and H. Jebelli. Worker-aware robotic motion planner in construction for improved psychological well-being during worker-robot interaction. In *Construction Research Congress 2022*, pages 205–214, 2022.

S. Lo Piano. Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, 7 (1):1–7, 2020.

E. Loizaga, A. T. Eyam, L. Bastida, and J. L. M. Lastra. A comprehensive study of human factors, sensory principles and commercial solutions for future human-centered working operations in Industry 5.0. *IEEE Access*, 2023.

R. Loomes, L. Hull, and W. P. L. Mandy. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6):466–474, 2017.

T. Loucas. Autism spectrum disorder. In *Supporting Young Children with Communication Problems*, pages 104–118. David Fulton Publishers, 2015.

H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 351–360, 2012.

L. Lu, Z. Xie, H. Wang, L. Li, and X. Xu. Mental stress and safety awareness during human-robot collaboration-review. *Applied ergonomics*, 105:103832, 2022a.

S. Lu, F. Wei, and G. Li. The evolution of the concept of stress and the framework of the stress system. *Cell stress*, 5(6):76, 2021.

Y. Lu, H. Zheng, S. Chand, W. Xia, Z. Liu, X. Xu, L. Wang, Z. Qin, and J. Bao. Outlook on human-centric manufacturing towards Industry 5.0. *Journal of Manufacturing Systems*, 62:612–627, 2022b.

P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Workshops*, pages 94–101. IEEE Computer Society, 2010.

P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 57–64. IEEE, 2011.

J. Luig and A. Sontacchi. A speech database for stress monitoring in the cockpit. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 228 (2):284–296, 2014.

M. A. Lumley, J. L. Cohen, G. S. Borszcz, A. Cano, A. M. Radcliffe, L. S. Porter, H. Schubiner, and F. J. Keefe. Pain and emotion: A biopsychosocial review of recent research. *Journal of clinical psychology*, 67(9):942–968, 2011.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

S. Luo and P. Johnston. A review of electrocardiogram filtering. *Journal of electrocardiology*, 43(6):486–496, 2010.

T. Luong, N. Martin, A. Raison, F. Argelaguet, J.-M. Diverrez, and A. Lécuyer. Towards real-time recognition of users mental workload using integrated physiological sensors into a VR HMD. In *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*, pages 425–437. IEEE, 2020.

S. Luqin. A survey of facial expression recognition based on convolutional neural network. In *18th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2019, Beijing, China, June 17-19, 2019*, pages 1–6. IEEE, 2019.

C. Lytridis, V. G. Kaburlasos, C. Bazinas, G. A. Papakostas, G. Sidiropoulos, V.-A. Nikopoulou, V. Holeva, M. Papadopoulou, and A. Evangeliou. Behavioral data analysis of robot-assisted autism spectrum disorder (ASD) interventions based on lattice computing techniques. *Sensors*, 22(2):621, 2022.

W. E. Mackay. DOIT: The design of interactive things. selected methods for quickly and effectively designing interactive systems from the user's perspective. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–3, 2023.

R. Mackinlay, T. Charman, and A. Karmiloff-Smith. High functioning children with autism spectrum disorder: A novel test of multitasking. *Brain and cognition*, 61(1):14–24, 2006.

R. Maeran and F. Cangiano. Flow experience and job characteristics: Analyzing the role of flow in job satisfaction. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, 20(1):13–26, 2013.

K. Magtibay and K. Umapathy. A review of tools and methods for detection, analysis, and prediction of allostatic load due to workplace stress. *IEEE Transactions on Affective Computing*, 2023.

B. Mahesh. Machine learning algorithms-A review. *International Journal of Science and Research (IJSR)*, 9(1):381–386, 2020.

M. H. Mahoor. AffectNet, 2017. URL `http://mohammadmahoor.com/affectnet/`. Accessed: 2023-11-27.

D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior research methods*, pages 1–8, 2021.

M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European heart journal*, 17(3):354–381, 1996.

B. A. Mammen, S. Irwin, and J. S. Tecklin. Common cardiac and pulmonary clinical measures. In *Cardiopulmonary Physical Therapy*, pages 177–224. Elsevier, 2004.

V. Markova, T. Ganchev, and K. Kalinkov. CLAS: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4. IEEE, 2019.

M. Martinho, A. Fred, and H. Silva. Towards continuous user recognition by exploring physiological multimodality: An electrocardiogram (ECG) and blood volume pulse (BVP) approach. In *2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI)*, pages 1–6. IEEE, 2018.

T. Masood and P. Sonntag. Industry 4.0: Adoption challenges and benefits for SMEs. *Computers in industry*, 121:103261, 2020.

F. Massimini, M. Csikszentmihalyi, and M. Carli. The monitoring of optimal experience: A tool for psychiatric rehabilitation. *Journal of Nervous and Mental Disease*, 175:545–549, 1987.

E. Matheson, R. Minto, E. G. Zampieri, M. Faccio, and G. Rosati. Human–robot collaboration in manufacturing applications: A review. *Robotics*, 8(4):100, 2019.

Y. Matsumoto, T. Ogasawara, and A. Zelinsky. Behavior recognition based on head pose and gaze direction measurement. In *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2127–2132. IEEE, 2000.

M. S. Matthias, A. T. Hirsh, S. Ofner, and J. Daggy. Exploring the relationships among social support, patient activation, and pain-related outcomes. *Pain Medicine*, 23(4):676–685, 2022.

B. J. Matuszewski, W. Quan, and L.-K. Shark. High-resolution comprehensive 3-D dynamic database for facial articulation analysis. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2128–2135. IEEE, 2011.

M. Mauri, P. Cipresso, A. Balgera, M. Villamira, and G. Riva. Why is facebook so successful? Psychophysiological measures describe a core flow state while using facebook. *Cyberpsychology, Behavior, and Social Networking*, 14(12):723–731, 2011.

P. Mayring. *Qualitative content analysis: Theoretical foundation, basic procedures and software solution.* AUT, Klagenfurt, 2014. URL https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173.

L. K. McCorry. Physiology of the autonomic nervous system. *American journal of pharmaceutical education*, 71(4), 2007.

K. T. McKay, S. A. Grainger, S. P. Coundouris, D. P. Skorich, L. H. Phillips, and J. D. Henry. Visual attentional orienting by eye gaze: A meta-analytic review of the gaze-cueing effect. *Psychological Bulletin*, 147(12):1269, 2021.

G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard, and E. André. Exploring a model of gaze for grounding in multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 247–254, 2014.

A. Mehrabian. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies.* Oelgeschlager, Gunn & Hain, Cambridge, 1980.

A. Mehrabian and J. A. Russell. *An approach to environmental psychology.* MIT Press, 1974.

A. Meissner, A. Trübswetter, A. S. Conti-Kufner, and J. Schmidtler. Friend or foe? Understanding assembly workers' acceptance of human-robot collaboration. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–30, 2020.

R. Melzack and P. D. Wall. Pain mechanisms: A new theory: A gate control system modulates sensory input from the skin before it evokes pain perception and response. *Science*, 150(3699):971–979, 1965.

P. Mende-Siedlecki, J. Qu-Lee, J. Lin, A. Drain, and A. Goharzad. The Delaware pain database: A set of painful expressions and corresponding norming data. *Pain reports*, 5(6):e853, 2020.

E. Mendoza and G. Carballo. Vocal tremor and psychological stress. *Journal of Voice*, 13(1): 105–112, 1999.

D. Mézière, L. Yu, E. D. Reichle, T. von der Malsburg, and G. M. McArthur. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 2021.

V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen, and D. Kotz. Evaluating the reproducibility of physiological stress detection models. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(4):1–29, 2020.

N. Mitev, P. Renner, T. Pfeiffer, and M. Staudte. Towards efficient human–machine collaboration: Effects of gaze-driven feedback and engagement on performance. *Cognitive Research: Principles and Implications*, 3(1):51, 2018.

S. Mittal, S. Mahendra, V. Sanap, and P. Churi. How can machine learning be used in stress management: A systematic literature review of applications in workplaces and education. *International Journal of Information Management Data Insights*, 2(2):100110, 2022.

A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

M. Mondellini, P. Prajod, M. L. Nicora, M. Chiappini, E. Micheletti, F. A. Storm, R. Vertechy, E. André, and M. Malosio. Behavioral patterns in robotic collaborative assembly: Comparing neurotypical and autism spectrum disorder participants. *Frontiers in Psychology*, 14, 2023.

M. Mondellini, M. L. Nicora, P. Prajod, E. André, R. Vertechy, A. Antonietti, and M. Malosio. Exploring the dynamics between cobot's production rhythm, locus of control and emotional state in a collaborative assembly scenario. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2024.

G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. Müller. Layer-wise relevance propagation: An overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 193–209. 2019.

A. Montoya, D. Holman, SF-data-science, T. Smith, and W. Kan. State farm distracted driver detection, 2016. URL `https://kaggle.com/competitions/state-farm-distracted-driver-detection`.

A. Moon, D. M. Troniak, B. Gleeson, M. K. Pan, M. Zheng, B. A. Blumer, K. MacLean, and E. A. Croft. Meet me where I'm gazing: How shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 334–341, 2014.

R. Mosberger and H. Andreasson. An inexpensive monocular vision system for tracking humans in industrial environments. In *2013 IEEE International Conference on Robotics and Automation*, pages 5850–5857. IEEE, 2013.

C. Mühlemeyer. Assessment and design of employees-cobot-interaction. In *Human Interaction and Emerging Technologies: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHIET 2019), August 22-24, 2019, Nice, France*, pages 771–776. Springer, 2020.

D. Mukherjee, K. Gupta, L. H. Chang, and H. Najjaran. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing*, 73:102231, 2022.

S. C. Müller and T. Fritz. Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 1, pages 688–699. IEEE, 2015.

M. Murin, J. Hellriegel, and W. Mandy. *Autism spectrum disorder and the transition into secondary school: A handbook for implementing strategies in the mainstream school setting*. Jessica Kingsley Publishers, 2016.

R. R. Murphy. *Introduction to AI robotics*. MIT press, 2019.

B. Mutlu, N. Roy, and S. Šabanović. Cognitive human–robot interaction. *Springer handbook of robotics*, pages 1907–1934, 2016.

L. Nacke and C. A. Lindley. Flow and immersion in first-person shooters: Measuring the player's gameplay experience. In *Proceedings of the 2008 conference on future play: Research, play, share*, pages 81–88, 2008.

F. F.-H. Nah, B. Eschenbrenner, Q. Zeng, V. R. Telaprolu, and S. Sepehr. Flow in gaming: Literature synthesis and framework development. *International Journal of Information Systems and Management*, 1(1-2):83–124, 2014.

J. Nakamura and M. Csikszentmihalyi. The concept of flow. *Handbook of positive psychology*, 89:105, 2002.

Y. Nakashima, J. Kim, S. Flutura, A. Seiderer, and E. André. Stress recognition in daily work. In *Pervasive Computing Paradigms for Mental Health: 5th International Conference, MindCare 2015, Milan, Italy, September 24-25, 2015, Revised Selected Papers 5*, pages 23–33. Springer, 2016.

M. Namvari, J. Lipoth, S. Knight, A. A. Jamali, M. Hedayati, R. J. Spiteri, and S. Syed-Abdul. Photoplethysmography enabled wearable devices and stress detection: A scoping review. *Journal of Personalized Medicine*, 12(11):1792, 2022.

V. Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4 (51-62):56, 2017.

S. Neupane, S. Mitra, I. A. Fernandez, S. Saha, S. Mittal, J. Chen, N. Pillai, and S. Rahimi. Security considerations in AI-Robotics: A survey of current methods, challenges, and opportunities. *IEEE Access*, 2024.

B. A. Newman, A. Biswas, S. Ahuja, S. Girdhar, K. K. Kitani, and H. Admoni. Examining the effects of anticipatory robot assistance on human decision making. In *International Conference on Social Robotics*, pages 590–603. Springer, 2020.

M. L. Nicora, E. André, D. Berkmans, C. Carissoli, T. D'Orazio, A. Delle Fave, P. Gebhard, R. Marani, R. M. Mira, L. Negri, et al. A human-driven control architecture for promoting good mental health in collaborative robot scenarios. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 285–291. IEEE, 2021.

M. L. Nicora, S. Beyrodt, D. Tsovaltzi, F. Nunnari, P. Gebhard, and M. Malosio. Towards social embodied cobots: The integration of an industrial cobot with a social virtual agent. *arXiv preprint arXiv:2301.06471*, 2023.

N. Nielsen Norman Group. Empathy mapping. `https://www.nngroup.com/articles/empathy-mapping/`, 2018. Accessed: 2024-03-07.

K. Nkurikiyeyezu, A. Yokokubo, and G. Lopez. The effect of person-specific biometrics in improving generic stress predictive models. *arXiv preprint arXiv:1910.01770*, 2019a.

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

C. Norsworthy, B. Jackson, and J. A. Dimmock. Advancing our understanding of psychological flow: A scoping review of conceptualizations, measurements, and applications. *Psychological bulletin*, 147(8):806, 2021.

F. Nunnari, M. L. Nicora, P. Prajod, S. Beyrodt, L. Chehayeb, E. André, P. Gebhard, M. Malosio, and D. Tsovaltzi. Understanding and mapping pleasure, arousal and dominance social signals to robot-avatar behavior. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2023.

T. Oka and S. Uchino. Human-robot cooperative conveyance using speech and head gaze. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 217–220, 2016.

R. Oliveira, P. Arriaga, P. Alves-Oliveira, F. Correia, S. Petisca, and A. Paiva. Friends or foes? Socioemotional support and gaze behaviors in mixed groups of humans and robots. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pages 279–288, 2018.

L. Onnasch and E. Roesler. A taxonomy to structure and analyze human–robot interaction. *International Journal of Social Robotics*, 13(4):833–849, 2021.

L. Onnasch, P. Schweidler, and M. Wieser. Effects of predictive robot eyes on trust and task performance in an industrial cooperation task. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 442–446, 2023.

K. O'Shea and R. Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

G. Osika. Humanistic services in the context of implementation Society 5.0. *Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie*, (183), 2023.

E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, and S. Walter. Cross-database evaluation of pain recognition from facial video. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 181–186. IEEE, 2019.

L. Paletta, M. Pszeida, H. Ganster, F. Fuhrmann, W. Weiss, S. Ladstätter, A. Dini, S. Murg, H. Mayer, I. Brijacak, et al. Gaze-based human factors measurements for the evaluation of intuitive human-robot collaboration in real-time. In *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1528–1531. IEEE, 2019.

O. Palinko, F. Rea, G. Sandini, and A. Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054. IEEE, 2016.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

J. A. C. Panceri, É. Freitas, J. C. de Souza, S. da Luz Schreider, E. Caldeira, and T. F. Bastos. A new socially assistive robot with integrated serious games for therapies with children with autism spectrum disorder and down syndrome: A pilot study. *Sensors*, 21(24):8414, 2021.

A. Papetti, F. Gregori, M. Pandolfi, M. Peruzzini, and M. Germani. A method to improve workers' well-being toward human-centered connected factories. *Journal of Computational Design and Engineering*, 7(5):630–643, 2020.

M. Parent, I. Albuquerque, A. Tiwari, R. Cassani, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk. PASS: A multimodal database of physical activity and stress for mobile passive body/brain-computer interface research. *Frontiers in Neuroscience*, 14:542934, 2020.

S. Park and R. Catrambone. Social facilitation effects of virtual humans. *Human factors*, 49 (6):1054–1060, 2007.

O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

S. K. Paul, M. Nicolescu, and M. Nicolescu. Integrating user gaze with verbal instruction to reliably estimate robotic task parameters in a human-robot collaborative environment. In *Proceedings of the 2023 6th International Conference on Machine Vision and Applications*, pages 51–58, 2023.

P. E. G. Paul Ekman Group. Universal emotions, 2023. URL `https://www.paulekman.com/universal-emotions/`. Accessed: 2024-05-16.

J. E. Peabody, R. Ryznar, M. T. Ziesmann, L. Gillman, R. J. Ryznar, and L. M. Gillman. A systematic review of heart rate variability as a measure of stress in medical professionals. *Cureus*, 15(1), 2023.

J. Pearce. Engaging the learner: How can the flow experience support E-Learning? In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 2288–2295. Association for the Advancement of Computing in Education (AACE), 2005.

L. Pecchia, R. Castaldo, L. Montesinos, and P. Melillo. Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations. *Healthcare technology letters*, 5(3):94–100, 2018.

H. Pedersen. Learning appearance features for pain detection using the UNBC-McMaster shoulder pain expression archive database. In *Computer Vision Systems: 10th International Conference, ICVS 2015, Copenhagen, Denmark, July 6-9, 2015, Proceedings 10*, pages 128–136. Springer, 2015.

C. Peifer. Psychophysiological correlates of flow-experience. *Advances in flow research*, pages 139–164, 2012.

C. Peifer and G. Wolters. Flow in the context of work. *Advances in flow research*, pages 287–321, 2021.

C. Peifer, A. Schulz, H. Schächinger, N. Baumann, and C. H. Antoni. The relation of flow-experience and physiological arousal under stress—Can u shape it? *Journal of Experimental Social Psychology*, 53:62–69, 2014.

C. Peifer, C. Syrek, V. Ostwald, E. Schuh, and C. H. Antoni. Thieves of flow: How unfinished tasks at work are related to flow experience and wellbeing. *Journal of Happiness Studies*, 21:1641–1660, 2020.

C. Peifer, G. Wolters, L. Harmat, J. Heutte, J. Tan, T. Freire, D. Tavares, C. Fonte, F. O. Andersen, J. van den Hout, et al. A scoping review of flow research. *Frontiers in Psychology*, 13:815665, 2022.

V. Pennazio, L. Fedeli, and E. Datteri. The use of robotics with children with ASD. Results from a pilot study on definition of target behaviors and prompts. In *INTED 2020 Proceedings*, pages 2147–2156. IATED, 2020.

P. Pennisi, A. Tonacci, G. Tartarisco, L. Billeci, L. Ruta, S. Gangemi, and G. Pioggia. Autism and social robotics: A systematic review. *Autism Research*, 9(2):165–183, 2016.

E. Peper, F. Shaffer, and I.-M. Lin. Garbage in; garbage out—Identify blood volume pulse (BVP) artifacts before analyzing and interpreting BVP, blood volume pulse amplitude, and heart rate/respiratory sinus arrhythmia data. *Biofeedback*, 38(1):19–23, 2010.

V. Petrolini, M. Jorba, and A. Vicente. What does it take to be rigid? Reflections on the notion of rigidity in autism. *Frontiers in Psychiatry*, 14:1072362, 2023.

U. J. Pfeiffer, K. Vogeley, and L. Schilbach. From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, 37(10):2516–2528, 2013.

T. Pham, Z. J. Lau, S. A. Chen, and D. Makowski. Heart rate variability in psychology: A review of HRV indices and an analysis tutorial. *Sensors*, 21(12):3998, 2021.

J. Piskorski and P. Guzik. Filtering poincare plots. *Computational methods in science and technology*, 11(1):39–48, 2005.

R. Plutchik. *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association, 2003.

E. Poljac, V. Hoofs, M. M. Princen, and E. Poljac. Understanding behavioural rigidity in autism spectrum conditions: The role of intentional control. *Journal of Autism and Developmental Disorders*, 47:714–727, 2017.

E. G. Popkova, Y. V. Ragulina, and A. V. Bogoviz. Fundamental differences of transition to Industry 4.0 from previous industrial revolutions. *Industry 4.0: Industrial revolution of the 21st century*, pages 21–29, 2019.

H. F. Posada-Quintero and K. H. Chon. Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors*, 20(2):479, 2020.

P. Prajod and E. André. On the generalizability of ECG-based stress detection models. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 549–554. IEEE, 2022.

P. Prajod, T. Huber, and E. André. Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions. In *International Conference on Multimedia Modeling*, pages 311–322. Springer, 2022a.

P. Prajod, D. Schiller, T. Huber, and E. André. Do deep neural networks forget facial action units?—Exploring the effects of transfer learning in health related facial expression recognition. *AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action*, 1013:217, 2022b.

P. Prajod, M. Lavit Nicora, M. Malosio, and E. André. Gaze-based attention recognition for human-robot collaboration. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 140–147, 2023a.

P. Prajod, M. L. Nicora, M. Mondellini, G. Tauro, R. Vertechy, M. Malosio, and E. André. Gaze detection and analysis for initiating joint activity in industrial human-robot collaboration. *arXiv preprint arXiv:2312.06643*, 2023b.

P. Prajod, M. Lavit Nicora, M. Mondellini, M. M. Falerni, R. Vertechy, M. Malosio, and E. André. Flow in human-robot collaboration—Multimodal analysis and perceived challenge detection in industrial scenarios. *Frontiers in Robotics and AI*, 11:1393795, 2024a.

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

P. Prajod, D. Schiller, D. W. Don, and E. André. Faces of experimental pain: Transferability of deep learned heat pain features to electrical pain. *arXiv preprint arXiv:2406.11808*, 2024c.

D. D. Price. Central neural mechanisms that interrelate sensory and affective dimensions of pain. *Molecular interventions*, 2(6):392, 2002.

K. M. Prkachin. The consistency of facial expressions of pain: A comparison across modalities. *Pain*, 51(3):297–306, 1992.

K. M. Prkachin. Assessing pain by facial expression: Facial expression as nexus. *Pain Research and Management*, 14:53–58, 2009.

P. D. Purnamasari, R. Martmis, and R. R. Wijaya. Stress detection application based on heart rate variability (HRV) and k-nearest neighbor (kNN). In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 271–276. IEEE, 2019.

M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng, et al. ROS: An open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

M. S. Rahman, A. Venkatachalapathy, A. Sharma, J. Wang, S. V. Gursoy, D. Anastasiu, and S. Wang. Synthetic distracted driving (SynDD1) dataset for analyzing distracted behaviors and various gaze zones of a driver. *Data in brief*, 46:108793, 2023.

S. N. Raja, D. B. Carr, M. Cohen, N. B. Finnerup, H. Flor, S. Gibson, F. J. Keefe, J. S. Mogil, M. Ringkamp, K. A. Sluka, et al. The revised international association for the study of pain definition of pain: Concepts, challenges, and compromises. *Pain*, 161(9):1976–1982, 2020.

A. Raptopoulou, A. Komnidis, P. D. Bamidis, and A. Astaras. Human–robot interaction for social skill development in children with ASD: A literature review. *Healthcare Technology Letters*, 8(4):90–96, 2021.

N. Rashid, L. Chen, M. Dautta, A. Jimenez, P. Tseng, and M. A. Al Faruque. Feature augmented hybrid CNN for stress recognition using wrist-based photoplethysmography sensor. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2374–2377. IEEE, 2021.

N. Rathee and D. Ganotra. Multiview distance metric learning on facial feature descriptors for automatic pain intensity detection. *Computer Vision and Image Understanding*, 147: 77–86, 2016.

D. F. Redaelli, F. A. Storm, and G. Fioretta. MindBot planetary gearbox, Nov. 2021. URL `https://doi.org/10.5281/zenodo.5675810`.

G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker. Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8: 54776–54788, 2020.

B. Reichard, F. Schrumpf, F. Anders, K. Bode, and M. Fuchs. Camera-based pain assessment during surgical interventions. In *Current Directions in Biomedical Engineering*, volume 8, pages 423–426. De Gruyter, 2022.

S. Rezaei, A. Moturu, S. Zhao, K. M. Prkachin, T. Hadjistavropoulos, and B. Taati. Unobtrusive pain monitoring in older adults with dementia using pairwise and contrastive training. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1450–1462, 2020.

F. Rheinberg. Die flow-kurzskala (FKS) übersetzt in verschiedene sprachen. The flow-short-scale (FSS) translated into various languages. 2015. doi: 10.13140/RG.2.1.4417.2243. URL `http://rgdoi.net/10.13140/RG.2.1.4417.2243`.

F. Rheinberg, R. Vollmeyer, and S. Engeser. Die erfassung des flow-erlebens. In *Diagnostik von Selbstkonzept, Lernmotivation und Selbstregulation [Diagnosis of Motivation and Self-Concept]*. Hogrefe, Göttingen, 2003.

J. L. Rhudy and M. W. Meagher. The role of emotion in pain modulation. *Current Opinion in Psychiatry*, 14(3):241–245, 2001.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

I. Rieger, R. Kollmann, B. Finzel, D. Seuss, and U. Schmid. Verifying deep learning-based decisions for facial expression recognition. *arXiv preprint arXiv:2003.00828*, 2020.

R. Rissler, M. Nadj, M. X. Li, N. Loewe, M. T. Knierim, and A. Maedche. To be or not to be in flow at work: Physiological classification of flow using machine learning. *IEEE Transactions on Affective Computing*, 2020.

S. Robla-Gómez, V. M. Becerra, J. R. Llata, E. Gonzalez-Sarabia, C. Torre-Ferrero, and J. Perez-Oria. Working together: A review on safe human-robot collaboration in industrial environments. *IEEE Access*, 5:26754–26773, 2017.

P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics*, 52(5):3314–3324, 2017.

M. R. Roller and P. J. Lavrakas. *Applied qualitative research design: A total quality framework approach.* Guilford Publications, 2015.

H. Romat, M.-A. Williams, X. Wang, B. Johnston, and H. Bard. Natural human-robot interaction using social cues. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 503–504. IEEE, 2016.

D. Romero, J. Stahre, T. Wuest, O. Noran, P. Bernus, Å. Fast-Berglund, and D. Gorecky. Towards an Operator 4.0 typology: A human-centric perspective on the fourth industrial revolution technologies. In *proceedings of the international conference on computers and industrial engineering (CIE46), Tianjin, China*, pages 29–31, 2016.

K. S. Rook. The functions of social bonds: Perspectives from research on social support, loneliness and social isolation. In *Social support: Theory, research and applications*, pages 243–267. Springer, 1985.

K. Rosander. Development of gaze control in early infancy. In *Oxford Research Encyclopedia of Psychology*. 2020.

F. Rossano, P. Brown, and S. C. Levinson. Gaze, questioning and culture. *Conversation analysis: Comparative perspectives*, 27, 2009.

R. N. Roy, N. Drougard, T. Gateau, F. Dehais, and C. P. Chanel. How can physiological computing benefit human-robot interaction? *Robotics*, 9(4):100, 2020.

S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh. An approach for automatic pain detection through facial expression. *Procedia Computer Science*, 84:99–106, 2016.

J. M. Rožanec, I. Novalija, P. Zajec, K. Kenda, H. Tavakoli Ghinani, S. Suh, E. Veliou, D. Papamartzivanos, T. Giannetsos, S. A. Menesidou, et al. Human-centric artificial intelligence architecture for Industry 5.0 applications. *International journal of production research*, 61 (20):6847–6872, 2023.

K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception. In *Computer graphics forum*, volume 34, pages 299–326. Wiley Online Library, 2015.

A. Ruiz, J. van de Weijer, and X. Binefa. Regularized multi-concept MIL for weakly-supervised facial behavior categorization. In *BMVC*, volume 7, page 8. Citeseer, 2014.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

M. H. Saad, M. I. Khalil, and H. M. Abbas. End-to-end driver distraction recognition using novel low lighting support dataset. In *2020 15th International Conference on Computer Engineering and Systems (ICCES)*, pages 1–6. IEEE, 2020.

R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang. UBFC-Phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14(1):622–636, 2021.

K. Sadeghniiat-Haghighi and Z. Yazdi. Fatigue management in the workplace. *Industrial psychiatry journal*, 24(1):12, 2015.

L. Y. Saltzman, T. C. Hansel, and P. S. Bordnick. Loneliness, isolation, and social support factors in post-COVID-19 mental health. *Psychological Trauma: Theory, Research, Practice, and Policy*, 12(S1):S55, 2020.

A. Saran, S. Majumdar, E. S. Short, A. Thomaz, and S. Niekum. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8615–8621. IEEE, 2018.

P. Sarkar and A. Etemad. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554, 2020.

A. Sauppé and B. Mutlu. The social impact of a robot co-worker in industrial settings. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3613–3622, 2015.

L. Scalera, S. Seriani, P. Gallina, M. Lentini, and A. Gasparetto. Human–robot interaction through eye tracking for artistic drawing. *Robotics*, 10(2):54, 2021.

B. Scassellati, H. Admoni, and M. Matarić. Robots for use in autism research. *Annual review of biomedical engineering*, 14:275–294, 2012.

B. R. Schadenberg, D. Reidsma, V. Evers, D. P. Davison, J. J. Li, D. K. Heylen, C. Neves, P. Alvito, J. Shen, M. Pantić, et al. Predictable robots for autistic children—Variance in robot behaviour, idiosyncrasies in autistic children's characteristics, and child–robot engagement. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(5):1–42, 2021.

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

P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven. Wearable-based affect recognition—A review. *Sensors*, 19(19):4079, 2019.

J. Schmidtler, V. Knott, C. Hölzel, and K. Bengler. Human centered assistance applications for the working environment of the future. *Occupational Ergonomics*, 12(3):83–95, 2015.

B. Schneider and R. Pea. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *Educational Media and Technology Yearbook: Volume 40*, pages 99–125, 2017.

R. Schwarzer and U. Schulz. Stressful life events. *Handbook of psychology*, 9:29–56, 2003.

R. Schwarzer and S. Taubert. Tenacious goal pursuits and striving toward personal growth: Proactive coping. In *Beyond coping: Meeting goals, visions and challenges*, pages 19–35, 2002.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

H. Selye. Stress and the general adaptation syndrome. *British medical journal*, 1(4667): 1383–1392, 1950.

H. Selye. Forty years of stress research: Principal remaining problems and misconceptions. *Canadian Medical Association Journal*, 115(1):53, 1976.

A. Semwal and N. D. Londhe. Head movement dynamics based pain detection using spatiotemporal network. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 204–209. IEEE, 2021a.

A. Semwal and N. D. Londhe. MVFNet: A multi-view fusion network for pain intensity assessment in unconstrained environment. *Biomedical Signal Processing and Control*, 67: 102537, 2021b.

A. Senju and G. Csibra. Gaze following in human infants depends on communicative signals. *Current biology*, 18(9):668–671, 2008.

A. Senju and M. H. Johnson. The eye contact effect: Mechanisms and development. *Trends in cognitive sciences*, 13(3):127–134, 2009.

C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine*, 14(2):410–417, 2009.

D. Seuss, A. Dieckmann, T. Hassan, J.-U. Garbas, J. H. Ellgring, M. Mortillaro, and K. Scherer. Emotion expression from different angles: A video database for facial expressions of actors shot by a camera array. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 35–41. IEEE, 2019.

F. Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:290215, 2017.

A. M. Shahsavarani, E. Azad Marz Abadi, and M. Hakimi Kalkhoran. Stress: Facts and theories through literature review. *International Journal of Medical Reviews*, 2(2):230–241, 2015.

M. Shaiqur Rahman, J. Wang, S. Velipasalar Gursoy, D. Anastasiu, S. Wang, and A. Sharma. Synthetic distracted driving (SynDD2) dataset for analyzing distracted behaviors and various gaze zones of a driver. *arXiv e-prints*, pages arXiv–2204, 2022.

A. Sharma and B. J. Singh. Evolution of industrial revolutions: A review. *International Journal of Innovative Technology and Exploring Engineering*, 9(11):66–73, 2020.

M. Sharma, A. Tomar, and A. Hazra. Edge computing for Industry 5.0: Fundamental, applications and research challenges. *IEEE Internet of Things Journal*, 2024.

K. A. Shaw, M. J. Maenner, J. Baio, A. Washington, D. L. Christensen, L. D. Wiggins, S. Pettygrove, J. G. Andrews, T. White, C. R. Rosenberg, et al. Early identification of autism spectrum disorder among children aged 4 years—Early autism and developmental disabilities monitoring network, six sites, united states, 2016. *MMWR Surveillance Summaries*, 69(3):1, 2020.

L. Shi, C. Copot, and S. Vanlanduit. Application of visual servoing and eye tracking glass in human robot interaction: A case study. In *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*, pages 515–520. IEEE, 2019.

L. Shi, C. Copot, and S. Vanlanduit. GazeEMD: Detecting visual intention in gaze-based human-robot interaction. *Robotics*, 10(2):68, 2021.

A. Shrivastava, V. M. Patel, J. K. Pillai, and R. Chellappa. Generalized dictionaries for multiple instance learning. *International Journal of Computer Vision*, 114:288–305, 2015.

Q. Shu, Q. Tu, and K. Wang. The impact of computer self-efficacy and technology dependence on computer-related technostress: A social cognitive theory perspective. *International Journal of Human-Computer Interaction*, 27(10):923–939, 2011.

K. Sikka, A. Dhall, and M. S. Bartlett. Classification and weakly supervised pain localization using multiple segment representation. *Image and vision computing*, 32(10):659–670, 2014.

A. C. Simões, A. Pinto, J. Santos, S. Pinheiro, and D. Romero. Designing human-robot collaboration (HRC) workspaces in industrial settings: A systematic literature review. *Journal of Manufacturing Systems*, 62:28–43, 2022.

D. Simon, K. D. Craig, F. Gosselin, P. Belin, and P. Rainville. Recognition and discrimination of prototypical dynamic expressions of pain and emotions. *PAIN®*, 135(1-2):55–64, 2008.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

A. K. Singh and S. Krishnan. ECG signal feature extraction trends in methods and applications. *BioMedical Engineering OnLine*, 22(1):22, 2023.

L. Sixt, M. Granz, and T. Landgraf. When explanations lie: Why many modified BP attributions fail. In *Proceedings of the 37th International Conference on Machine Learning ICML 2020*, pages 9046–9057, 2020.

R. Sjouwerman and T. Lonsdorf. Latency of skin conductance responses across stimulus modalities. *Psychophysiology*, 56(4):e13307, 2019.

E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine*, 1(1):67, 2018.

B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280, 2013.

V. Srinivasan and R. Murphy. A survey of social gaze. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 253–254, 2011.

S. Sriramprakash, V. D. Prasanna, and O. R. Murthy. Stress detection in working people. *Procedia computer science*, 115:359–366, 2017.

A. Stahelski, A. Anderson, N. Browitt, and M. Radeke. Facial expressions and emotion labels are separate initiators of trait inferences from the face. *Frontiers in Psychology*, 12: 749933, 2021.

F. Stamatelopoulou, C. Pezirkianidis, E. Karakasidou, A. Lakioti, and A. Stalikas. "Being in the zone": A systematic review on the relationship of psychological correlates and the occurrence of flow experiences in sports' performance. *Psychology*, 9(08):2011, 2018.

C. J. Stanton and C. J. Stevens. Don't stare at me: The impact of a humanoid robot's gaze upon trust during a cooperative human–robot visual task. *International Journal of Social Robotics*, 9:745–753, 2017.

R. Sterna, P. Strojny, and K. Rębilas. Can virtual observers affect our behavior? *Social Psychological Bulletin*, 14(3):1–18, 2019.

F. A. Storm, M. Chiappini, C. Dei, C. Piazza, E. André, N. Reißner, I. Brdar, A. Delle Fave, P. Gebhard, M. Malosio, et al. Physical and mental well-being of cobot workers: A scoping review using the Software-Hardware-Environment-Liveware-Liveware-Organization model. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 2022.

K. Strabala, M. K. Lee, A. Dragan, J. Forlizzi, S. S. Srinivasa, M. Cakmak, and V. Micelli. Toward seamless human-robot handovers. *Journal of Human-Robot Interaction*, 2(1): 112–132, 2013.

D. Strazdas, J. Hintz, A.-M. Felßberg, and A. Al-Hamadi. Robots and wizards: An investigation into natural human–robot interaction. *IEEE Access*, 8:207635–207642, 2020.

J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

B. Su, S. Jung, L. Lu, H. Wang, L. Qing, and X. Xu. Exploring the impact of human-robot interaction on workers' mental stress in collaborative assembly tasks. *Applied Ergonomics*, 116:104224, 2024.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis. A multimodal dataset for various forms of distracted driving. *Scientific data*, 4(1):1–21, 2017.

J. Tao and T. Tan. Affective computing: A review. In *International Conference on Affective Computing and Intelligent Interaction*, pages 981–995. Springer, 2005.

M. Tarabini, M. Marinoni, M. Mascetti, P. Marzaroli, F. Corti, H. Giberti, A. Villa, and P. Mascagni. Monitoring the human posture in industrial environment: A feasibility study. In *2018 IEEE sensors applications symposium (SAS)*, pages 1–6. IEEE, 2018.

P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak. Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184, 2017.

M. Tavakolian, M. B. Lopez, and L. Liu. Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. *Pattern Recognition Letters*, 140: 26–33, 2020.

I. R. Tayibnapis, M.-K. Choi, and S. Kwon. Driver's gaze zone estimation by transfer learning. In *2018 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–5. IEEE, 2018.

P. Taylor, N. Griffiths, A. Bhalerao, Z. Xu, A. Gelencser, and T. Popham. Warwick-JLR driver monitoring dataset (DMD) statistics and early findings. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 89–92, 2015.

Y. Terzioğlu, B. Mutlu, and E. Şahin. Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 343–357, 2020.

L. Thau, V. Reddy, and P. Singh. Anatomy, central nervous system. In *StatPearls*. StatPearls Publishing, 2022.

P. Thiam, H. A. Kestler, and F. Schwenker. Two-stream attention network for pain recognition from video sequences. *Sensors*, 20(3):839, 2020.

Y. Tian, Y. Bian, P. Han, P. Wang, F. Gao, and Y. Chen. Physiological signal analysis for evaluating flow during playing of computer games of varying difficulty. *Frontiers in Psychology*, 8:1121, 2017.

Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

A. Toichoa Eyam, W. M. Mohammed, and J. L. Martinez Lastra. Emotion-driven analysis and control of human-robot interactions in collaborative applications. *Sensors*, 21(14): 4626, 2021.

A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.

M. T. Tomczak. Employees with autism spectrum disorders in the digitized work environment: Perspectives for the future. *Journal of disability policy studies*, 31(4):195–205, 2021.

Y. Topoglu, J. Watson, R. Suri, and H. Ayaz. Electrodermal activity in ambulatory settings: A narrative review of literature. In *Advances in Neuroergonomics and Cognitive Engineering. AHFE 2019*, pages 91–102. Springer, 2020.

T. Tozman, E. S. Magdas, H. G. MacDougall, and R. Vollmeyer. Understanding the psychophysiology of flow: A driving simulator experiment to investigate the relationship between flow and heart rate variability. *Computers in Human Behavior*, 52:408–418, 2015.

T.-A. Tran, J. Abonyi, L. Kovács, G. Eigner, and T. Ruppert. Heart rate variability measurement to assess work-related stress of physical workers in manufacturing industries - Protocol for a systematic literature review. In *2022 IEEE 20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, pages 313–318. IEEE, 2022.

C. Tronstad, M. Amini, D. R. Bach, and Ø. G. Martinsen. Current trends and opportunities in the methodology of electrodermal activity measurement. *Physiological measurement*, 43(2):02TR01, 2022.

D. C. Turk and H. Flor. Etiological theories and treatments for chronic back pain. ii. psychological models and interventions. *Pain*, 19(3):209–233, 1984.

M. Umair, N. Chalabianloo, C. Sas, and C. Ersoy. HRV and stress: A mixed-methods approach for comparison of wearable heart rate sensors for biofeedback. *IEEE Access*, 9: 14005–14024, 2021.

S. Upasani, D. Srinivasan, Q. Zhu, J. Du, and A. Leonessa. Eye-tracking in physical human–robot interaction: Mental workload and performance prediction. *Human factors*, page 00187208231204704, 2023.

M. Valstar, S. Zafeiriou, and M. Pantic. Facial actions as social signals. In *Social signal processing*. Cambridge University Press, 2017.

B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Meudt, H. Neumann, J. Kim, et al. The SenseEmotion database: A multimodal database for the development and systematic validation of an automatic pain-and emotion-recognition system. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction: 4th IAPR Workshop*, pages 127–139. Springer, 2017.

P. E. Velmovitsky, P. Alencar, S. T. Leatherdale, D. Cowan, and P. P. Morita. Towards real-time public health: A novel mobile health monitoring system. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 6049–6051. IEEE, 2021.

E. Verna, S. Puttero, G. Genta, and M. Galetto. A novel diagnostic tool for human-centric quality monitoring in human-robot collaboration manufacturing. *Journal of Manufacturing Science and Engineering*, 145(12):121009, 2023.

V. Villani, F. Pini, F. Leali, and C. Secchi. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, 55:248–266, 2018.

C. L. von Baeyer, T. Piira, C. T. Chambers, M. Trapanotto, and L. K. Zeltzer. Guidelines for the cold pressor task as an experimental pain stimulus for use with children. *The journal of Pain*, 6(4):218–227, 2005.

A. M. von der Pütten, N. C. Krämer, J. Gratch, and S.-H. Kang. "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior*, 26 (6):1641–1650, 2010.

S. Vora, A. Rangesh, and M. M. Trivedi. On generalizing driver gaze zone estimation using convolutional neural networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 849–854. IEEE, 2017.

G. Vos, K. Trinh, Z. Sarnyai, and M. R. Azghadi. Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices. *Journal of Biomedical Informatics*, 148:104556, 2023a.

G. Vos, K. Trinh, Z. Sarnyai, and M. R. Azghadi. Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. *International Journal of Medical Informatics*, 173:105026, 2023b.

R. S. Wadhwa. Flexibility in manufacturing automation: A living lab case study of norwegian metalcasting SMEs. *Journal of Manufacturing Systems*, 31(4):444–454, 2012.

J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The social signal interpretation (SSI) framework: Multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 831–834, 2013.

S. Walter and A. Al-Hamadi. The BioVid heat pain database, 2022. URL `https://www.nit.ovgu.de/BioVid.html`. Accessed: 2023-12-12.

S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. M. da Silva. The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE international conference on cybernetics (CYBCO)*, pages 128–131. IEEE, 2013.

F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille. Regularizing face verification nets for pain intensity regression. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1087–1091. IEEE, 2017.

J. Wang, W. Li, F. Li, J. Zhang, Z. Wu, Z. Zhong, and N. Sebe. 100-driver: A large-scale, diverse dataset for distracted driver classification. *IEEE Transactions on Intelligent Transportation Systems*, 24(7):7061–7072, 2023.

N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill. Is it my looks? Or something I said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *Persuasive Technology: 13th International Conference, PERSUASIVE 2018*, pages 56–69. Springer, 2018.

Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022.

J. A. Waxenbaum, V. Reddy, and M. Varacallo. *Anatomy, Autonomic Nervous System.* StatPearls Publishing, 2023.

D. Weber, W. Fuhl, E. Kasneci, and A. Zell. Multiperspective teaching of unknown objects via shared-gaze-based multimodal human-robot interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 544–553, 2023.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas. Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. *tm-Technisches Messen*, 86(7-8):404–412, 2019.

K. S. Welfare, M. R. Hallowell, J. A. Shah, and L. D. Riek. Consider the human work experience when integrating robotics in the workplace. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 75–84. IEEE, 2019.

P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, 8(3):286–299, 2016.

P. Werner, A. Al-Hamadi, and S. Walter. Analysis of facial expressiveness during experimentally induced heat pain. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 176–180. IEEE, 2017.

B. Wheaton and S. Montazer. Stressors, stress, and distress. *A handbook for the study of mental health: Social contexts, theories, and systems*, 2:171–199, 2010.

K. Wiech and I. Tracey. The influence of negative emotions on pain: Behavioral effects and neural mechanisms. *Neuroimage*, 47(3):987–994, 2009.

A. C. d. C. Williams. Facial expression of pain: An evolutionary account. *Behavioral and brain sciences*, 25(4):439–455, 2002.

A. C. d. C. Williams. Facial expressions of pain: Clinical meaning and research possibilities. *Pain Management*, 1(4):303–305, 2011.

A. C. d. C. Williams and K. D. Craig. Updating the definition of pain. *Pain*, 157(11):2420–2423, 2016.

E. A. Witt, J. Kenworthy, G. Isherwood, and W. C. Dunlop. Examining the association between pain severity and quality-of-life, work-productivity loss, and healthcare resource use among european adults diagnosed with pain. *Journal of Medical Economics*, 19(9): 858–865, 2016.

C. Y. Wong, L. Vergez, and W. Suleiman. Vision-and tactile-based continuous multimodal intention and attention recognition for safer physical human–robot interaction. *IEEE Transactions on Automation Science and Engineering*, 2023.

C. Wu, S. Wang, and Q. Ji. Multi-instance hidden markov model for facial expression recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–6. IEEE, 2015.

Y. Wu and Q. Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019.

D. Wurhofer, T. Meneweger, V. Fuchsberger, and M. Tscheligi. Deploying robots in a production environment: A study on temporal transitions of workers' experiences. In *Human-Computer Interaction–INTERACT 2015*, pages 203–220. Springer, 2015.

V.-R. Xefteris, M. Dominguez, J. Grivolla, A. Tsanousa, F. Zaffanela, M. Monego, S. Symeonidis, S. Diplaris, L. Wanner, S. Vrochidis, et al. A multimodal late fusion framework for physiological sensor and audio-signal-based stress detection: An experimental study and public dataset. *Electronics*, 12(23):4871, 2023.

X. Xiang, F. Wang, Y. Tan, and A. L. Yuille. Imbalanced regression for intensity series of pain expression from videos by regularizing spatio-temporal face nets. *Pattern Recognition Letters*, 163:152–158, 2022.

B. H. Ximenes and G. L. Ramalho. Concrete ethical guidelines and best practices in machine learning development. In *2021 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–8. IEEE, 2021.

L. D. Xu, E. L. Xu, and L. Li. Industry 4.0: State of the art and future trends. *International journal of production research*, 56(8):2941–2962, 2018.

X. Xu, J. S. Huang, and V. R. de Sa. Pain evaluation in video using extended multitask learning from multidimensional measurements. In *ML4H@ NeurIPS*, pages 141–154, 2019.

X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang. Industry 4.0 and Industry 5.0—Inception, conception and perception. *Journal of manufacturing systems*, 61:530–535, 2021.

K. Yamada, K. Matsudaira, H. Imano, A. Kitamura, and H. Iso. Influence of work-related psychosocial factors on the prevalence of chronic pain and quality of life in patients with chronic pain. *BMJ open*, 6(4):e010356, 2016.

T. Yamashita and S. N. Kudoh. Elucidation of validity of emotion model on EEG and facial expression. In *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pages 1–4. IEEE, 2022.

J. Yan, G. Lu, X. Li, W. Zheng, C. Huang, Z. Cui, Y. Zong, M. Chen, Q. Hao, Y. Liu, et al. FENP: A database of neonatal facial expression for pain analysis. *IEEE Transactions on Affective Computing*, 14(1):245–254, 2020.

R. Yang, S. Tong, M. Bordallo, E. Boutellaa, J. Peng, X. Feng, and A. Hadid. On pain assessment from facial videos using spatio-temporal local descriptors. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2016.

T.-X. Yang, W. Xie, C.-S. Chen, M. Altgassen, Y. Wang, E. F. Cheung, and R. C. Chan. The development of multitasking in children aged 7–12 years: Evidence from cross-sectional and longitudinal data. *Journal of experimental child psychology*, 161:63–80, 2017.

C.-T. Yen and K.-H. Li. Discussions of different deep transfer learning models for emotion recognitions. *IEEE Access*, 10:102860–102875, 2022.

J. A. Yeow, P. K. Ng, K. S. Tan, T. S. Chin, and W. Y. Lim. Effects of stress, repetition, fatigue and work environment on human error in manufacturing industries. *Journal of Applied Sciences*, 14(24):3464–3471, 2014.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

B. Yu, M. Funk, J. Hu, Q. Wang, and L. Feijs. Biofeedback for everyday stress management: A systematic review. *Frontiers in ICT*, 5:23, 2018.

Y. Yu and J.-M. Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020.

X. Yuan, S. Zhang, C. Zhao, X. He, B. Ouyang, and S. Yang. Pain intensity recognition from masked facial expressions using swin-transformer. In *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 723–728. IEEE, 2022.

Z. Zakeri, A. Arif, A. Omurtag, P. Breedon, and A. Khalid. Multimodal assessment of cognitive workload using neural, subjective and behavioural measures in smart factory settings. *Sensors*, 23(21):8926, 2023.

S. Zambon. *From Industry 4.0 to Society 5.0: Digital manufacturing technologies and the role of workers*. Università degli studi di Padova, 2022. Masters Thesis.

A. Zamkah, T. Hui, S. Andrews, N. Dey, F. Shi, and R. S. Sherratt. Identification of suitable biomarkers for stress and emotion detection for future personal affective wearable sensors. *Biosensors*, 10(4):40, 2020.

G. Zamzmi, R. Paul, D. Goldgof, R. Kasturi, and Y. Sun. Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN). In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

P. Zaparas, P. Katranitsiotis, K. Stavridis, and P. Daras. Detecting human distraction from gaze: An augmented reality approach in the robotic environment. In *Conference on Biomimetic and Biohybrid Systems*, pages 56–62. Springer, 2023.

Y. Zarghami, S. Mafeld, A. Conway, and B. Taati. Pain detection in masked faces during procedural sedation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2023.

M. R. S. Zawad, C. S. A. Rony, M. Y. Haque, M. H. A. Banna, M. Mahmud, and M. S. Kaiser. A hybrid approach for stress prediction from heart rate variability. In *Frontiers of ICT in Healthcare: Proceedings of EAIT 2022*, pages 111–121. Springer, 2023.

C. Zhang, Z. Wang, G. Zhou, F. Chang, D. Ma, Y. Jing, W. Cheng, K. Ding, and D. Zhao. Towards new-generation human-centric smart manufacturing in Industry 5.0: A systematic review. *Advanced Engineering Informatics*, 57:102121, 2023a.

P. Zhang, F. Li, R. Zhao, R. Zhou, L. Du, Z. Zhao, X. Chen, and Z. Fang. Real-time psychological stress detection according to ECG using deep learning. *Applied Sciences*, 11(9): 3838, 2021.

X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. ETH-Xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference*, pages 365–381. Springer, 2020.

X. Zhang, X. Wei, Z. Zhou, Q. Zhao, S. Zhang, Y. Yang, R. Li, and B. Hu. Dynamic alignment and fusion of multimodal physiological patterns for stress recognition. *IEEE Transactions on Affective Computing*, 2023b.

Y. Zhang, K. Pfeuffer, M. K. Chong, J. Alexander, A. Bulling, and H. Gellersen. Look together: Using gaze for assisting co-located collaborative search. *Personal and Ubiquitous Computing*, 21:173–186, 2017.

L. Zhao, X. Niu, L. Wang, J. Niu, X. Zhu, and Z. Dai. Stress detection via multimodal multi-temporal-scale fusion: A hybrid of deep learning and handcrafted feature approach. *IEEE Sensors Journal*, 2023.

R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3466–3474, 2016.

R. Zhi and M. Wan. Dynamic facial expression feature learning based on sparse RNN. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 1373–1377. IEEE, 2019.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.