Deep Learning Approaches for User Engagement Detection in Human-Robot Interaction: A Scoping Review

Downloaded from: https://research.chalmers.se, 2025-11-13 08:57 UTC

Citation for the original published paper (version of record):
Ravandi, B., Khan, I., Gander, P. et al (2025). Deep Learning Approaches for User Engagement Detection in Human-Robot Interaction: A Scoping
Review. International Journal of Human-Computer Interaction, 41(20): 13074-13092.
http://dx.doi.org/10.1080/10447318.2025.2470277

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Deep Learning Approaches for User Engagement Detection in Human-Robot Interaction: A Scoping Review

Bahram Salamat Ravandi, Imran Khan, Pierre Gander & Robert Lowe

To cite this article: Bahram Salamat Ravandi, Imran Khan, Pierre Gander & Robert Lowe (2025) Deep Learning Approaches for User Engagement Detection in Human-Robot Interaction: A Scoping Review, International Journal of Human–Computer Interaction, 41:20, 13074-13092, DOI: 10.1080/10447318.2025.2470277

To link to this article: https://doi.org/10.1080/10447318.2025.2470277

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 06 Mar 2025.

Submit your article to this journal

Article views: 1538

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=hihc20

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2025, VOL. 41, NO. 20, 13074–13092
https://doi.org/10.1080/10447318.2025.2470277

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Deep Learning Approaches for User Engagement Detection in Human-Robot Interaction: A Scoping Review

Bahram Salamat Ravandi, Imran Khan, Pierre Gander, and Robert Lowe

Department of Applied IT, University of Gothenburg, Gothenburg, Sweden

ABSTRACT

The increasing use of social assistive robots (SARs) has sparked researcher interest to investigate user engagement to enhance SAR interactive capabilities. Engagement in Human-Robot Interaction (HRI) aims to benefit users during interactions. Diverse interpretations of engagement have led to various metrics for its measurement and detection. Despite numerous algorithmic approaches for detecting user engagement, Deep Learning (DL) algorithms have become prominent in HRI engagement detection. However, there is a lack of comprehensive reviews on DL methods for engagement detection in HRI. This scoping review summarizes a decade of DL applications in HRI engagement detection, highlighting key findings and gaps including the need for context-specific datasets, understanding temporal dynamics, and exploring non-social robots. Moreover, this review focuses on employed DL algorithms, sensory inputs, ground truths, robots, and datasets. This review serves as a valuable reference for HRI researchers aiming to improve user engagement detection strategies.

KEYWORDS

Engagement; Social Robots; Deep Learning; Affective Computing; Human-Robot Interaction

1. Introduction

The field of Human-Robot Interaction (HRI) has seen significant development in recent years, finding applications in entertainment (Lytridis et al., 2019), healthcare (Weng & Hirata, 2022), and assistive technology (Kubota et al., 2022). Over the past two decades, a new type of robot called Socially Assistive Robots (SARs) has been developed which can assist human users through social interaction, aiming to achieve measurable progress in areas such as convalescence, rehabilitation, learning, and other assistive tasks (Feil-Seifer & Mataric, 2005). SARs focus on creating close and effective interactions with users to offer support and aid in various domains. These robots can be used to impact users' engagement through two-way interactions, such as providing verbal hints when playing a game (Jain et al., 2020), or displays of affective expressions, such as smiling or gesturing (Ritschel et al., 2017).

Research has shown that people's level of motivation and productivity can significantly increase in the presence of others. This phenomenon is often referred to as social facilitation, where the mere presence of others can enhance an individual's performance on a given task (Belletier et al., 2019). It's not just the presence of humans that can have the social presence effect, however; recent studies have also demonstrated that the presence of social robots can provide similar benefits (Luria et al., 2019). Social robots through adaptive social feedback (e.g., audiovisual feedback and/or affective interactions) have been found to improve task engagement and attention (Chan & Nejat, 2010), learning (Silvera-Tawil et al., 2022), and engagement (Ahmad et al., 2019). and attitudes to therapies (Logan et al., 2019).

The concept of "engagement" in the HRI literature is broad and encompasses various interpretations such as cognitive engagement, behavioral engagement, and emotional engagement, according to Doherty and Doherty (2019). The various interpretations of engagement in the HRI literature have therefore led to different features and measurements being used to evaluate user engagement. For example (Hadfield et al., 2019), used gaze and pose detection to detect engagement while (Mollahosseini et al., 2018) employed facial expressions to detect engagement. Moreover, considering the distinctions between different types of interactions, researchers can address diverse aspects of engagement, such as social behaviors, communication dynamics, and user preferences. For instance, the assessment of engagement can be conducted either online (i.e., in real-time applications) or offline (i.e., after the interaction). In studies that attempt to measure engagement online (algorithmic), the aim is to modify the interaction or provide feedback; where engagement is assessed offline (non-algorithmic) the design of the interaction is typically emphasized. As an example of an offline approach (Mucchiani et al., 2021), utilized surveys to measure users'

CONTACT Bahram Salamat Ravandi bahramsalamat@ait.gu.se Department of Applied IT, University of Gothenburg, Gothenburg, Sweden

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

responses, both verbal and physical, in order to determine which aspects of HRI are significant in the development of social agents for patient screening. To classify engagement effectively, various learning methods (Amaro et al., 2019; Andriella et al., 2020; Bärenholdt et al., 2020; Björling et al., 2018; Liu et al., 2007) have been utilized for the classification of engagement in both online and offline applications. However, the diversity of interpretations presents challenges for selecting optimal methodological approaches to enhance HRI frameworks, including the identification of suitable sensory data.

To clarify our definition of engagement, in this scoping review paper we incorporated two of the most common views in the literature (Doherty & Doherty, 2019):

1. "By engagement, we mean the process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in the interaction, evaluating whether to stay involved, and deciding when to end the connection," (Sidner et al., 2004, p. 78).
2. "Engagement is a quality of user experiences with technology that is characterized by challenge, aesthetic and sensory, appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, interest, and affect." (O'Brien & Toms, 2008, p. 949).

The first definition aids in understanding the establishment of engagement, while the second is geared towards assessing the quality of engagement. Various sensory inputs/features are employed to assess both the existence and quality of engagement. For instance, distance from the robot, and gaze towards/away from the robot serve as effective measures for determining the presence and level of engagement, and affect measures become more relevant in evaluating the depth or intensity of engagement.

In recent years, the use of machine learning techniques for detecting engagement has become increasingly prevalent (Oertel et al., 2020). Our particular focus pertains to Deep Learning (DL) methodologies, which have garnered attention for their inherent capacity to deal with large and/or unstructured multi-modal data through performing feature extraction and modality fusion (Goodfellow et al., 2016), without relying on pre-trained modules (e.g., Sanghvi et al., 2011), used the CAMShift algorithm from OpenCV packages to extract Body Lean Angle (BLA) (Salam et al., 2017), used different frontal and profile face detection models based on the OpenCV version of Viola and Jones Haar Cascade algorithm (Boccanfuso et al., 2016), used Principal Component Analysis (PCA) for dimensional reduction of data). Deep neural networks, such as CNNs and RNNs, offer the capability to extract complex features from data and leverage temporal information for more accurate estimations in HRI scenarios (Bandi & Thomas, 2023; Dhaussy et al., 2023; Gonzalez & Mizuuchi et al., 2023; Kawahara et al., 2018; Kokate et al., 2022; Shenoy et al., 2022; Simões et al., 2023; Świetlicka et al., 2023; Zhang et al., 2021).

Although there have been reviews on the synthesis of information relevant to assessing engagement, none, to the authors' knowledge, have explicitly delved into a systematic or scoping review of the DL methods utilized for engagement detection or discussed the sensory input/features employed for engagement detection. As an example of a non-systematic/scoping review (Doherty & Doherty, 2019), provides valuable insights into the concept of engagement in HCI research and mainly focuses on the concept of engagement, theoretical frameworks, and measurements. Another example by Oertel et al. (2020) aimed to answer several questions related to engagement, such as the most commonly used definitions of engagement, how engagement differs across different interaction settings and user types, and what methods are used to establish engagement ground truths. They suggest that choosing appropriate input features is crucial for successful engagement recognition using DL approaches. Notwithstanding the insights these papers provide, there is a lack of review articles summarizing how different data modalities or features such as speech features, facial features, and physiological sensors are integrated for engagement detection using DL methods, which features are commonly utilized, and how these features correspond with specific DL algorithms employed in modeling engagement, and how the ground truth for engagement detection is established.

In order to gain a comprehensive understanding of the existing applications of DL methods in engagement detection within the domain of HRI, this paper presents a scoping review to address several research questions. These questions aim to identify existing research gaps and provide researchers with an understanding of the current state of the literature with respect to the development of DL frameworks for engagement detection. These questions are designed to assist researchers in devising effective strategies for studying engagement in potential HRI setups and understanding commonly utilized features, thereby enabling them to equip their setups with appropriate sensors and engagement detection modules. Additionally, the review seeks to address challenges associated with the ground truths establishment of engagement, which often involves complexities and may require trained human annotators or alternative methods such as self-reporting. The review also explores the utilization of datasets from previous studies and examines engagement detection methodologies employed with specific social robots. Lastly, the review investigates methods for integrating engagement detection into HRI setups. The research questions guiding this review are as follows:

1. Which DL methods are used for users' engagement modeling in HRI?
2. Which sensory inputs/features are used for users' engagement modeling using DL methods in HRI?
3. How does the literature address the establishment of ground truths for engagement in HRI for training DL models?
4. What datasets have been used for training DL models for engagement detection in HRI?
5. What types of HRI setups, such as social robots, virtual agents, or interactive environments, are used to develop engagement detection models?

6. What engagement adaptation methods have been applied to the HRI setups?

The novelty of this work lies in its focused exploration of multimodal data integration and the application of advanced DL techniques for engagement detection in HRI. While previous studies have addressed engagement detection, this research uniquely emphasizes the synergistic use of diverse data modalities—such as facial features, speech features, and physiological signals—to create a more comprehensive understanding of user engagement. Furthermore, the work highlights the implementation of state-of-the-art DL algorithms, specifically tailored for processing and integrating multimodal data.

2. Methodology

The methodology employed in this review entailed a multi-step screening process. Our methodology was inspired by similar works previously published in HRI (Amirova et al., 2021; Oertel et al., 2020). In order to ensure unbiased assessments and minimize potential biases, we followed a rigorous approach that involved implementing inter-rater evaluations with blinding procedures in place. All four authors took part in the inter-rater screening, with each reviewer conducting screening and reviews independently, without knowledge of their peers' evaluations. In the event of conflicts arising, a conflict resolution process was conducted by an independent reviewer (not involved in the screening of the specific paper) to ensure that the inclusion or exclusion of papers was based on a majority vote decision.

Each paper was reviewed by two independent reviewers. Inter-rater reliability was assessed using Cohen's Kappa (κ), which measures agreement beyond chance (McHugh, 2012). For the title and abstract screening stage, the average κ value was 0.49, indicating moderate agreement. In the full-text screening stage, the average κ value was 0.63, reflecting substantial agreement.

2.1. Search

We conducted a literature search in four major databases—Scopus, IEEE Xplore, PubMed, and Elsevier—chosen for their extensive coverage of peer-reviewed research in robotics, computer science, and HRI. To supplement our search, we also applied backward snowballing to identify additional relevant studies and and set no start year limit. This search occurred initially in September 2022, with an updated search conducted in February 2024. In order to not overlook potential papers in the initial literature survey, we employed a search string that we hoped would capture a wide range of papers, to remove irrelevant papers in future stages of the process. Below are the detailed search queries used for each database:

- **Elsevier (via ScienceDirect):** `Title, abstract, keywords: Engagement AND (Human-Robot Interaction OR Virtual Agents OR HRI OR Human Robot Interaction OR Human-Robot).`
- **Scopus:** `TITLE-ABS("Engagement") AND TITLE-ABS("Human-Robot Interaction" OR "Virtual Agents" OR "HRI" OR "Human Robot Interaction" OR "Human-Robot").`
- **PubMed:** `("Engagement"[Title/Abstract]) AND (("Human-Robot Interaction"[Title/Abstract] OR "Virtual Agents"[Title/Abstract] OR "HRI"[Title/Abstract] OR "Human Robot Interaction"[Title/Abstract] OR "Human-Robot"[Title/Abstract]).`
- **IEEE Xplore:** `("Abstract":Engagement) AND ("Abstract":Human-Robot Interaction OR "Abstract":Virtual Agents OR "Abstract":HRI OR "Abstract":Human Robot Interaction OR "Abstract":Human-Robot).`

We decided to include "Virtual Agents" as a keyword in order to ensure a comprehensive review by capturing studies with overlapping methodologies and insights. This approach broadens the scope, accommodates terminology variations, and includes relevant algorithmic models for engagement that apply to both virtual agents and robots. Moreover, it is important to note that this study did not aim at restricting to a specific definition of "engagement," but instead focused on conducting a comprehensive search for research studies implementing algorithmic models for engagement detection in HRI, with a focus on DL methods.

2.2. Inclusion and exclusion criteria

In this study, we specifically excluded papers that did not focus on algorithmic detection of engagement. The inclusion criteria for papers in this study were as follows: the paper must present specific engagement algorithms that have undergone some training and are not limited to conceptual approaches. The paper must be written in legible English. Additionally, review papers such as scoping reviews, systematic reviews, meta-analyses, reviews of proceeding conferences, and publications of low quality such as non-peer-reviewed articles, book chapters, and editorials were excluded from this study.

Inclusion Criteria

- Application or investigation of engagement detection algorithms.
- Includes social robots, virtual agents, or HRI contexts.
- Simulated or physical robots are acceptable.
- Wizard of Oz (WoZ) studies if engagement data is collected for deep learning applications.
- Written in legible English.
- Peer-reviewed papers, including conference papers (e.g., IEEE, ICRA, IROS).
- Sufficient detail for replication or empirical investigation.
- Papers presenting trained algorithms (not just conceptual).

Exclusion Criteria

- Engagement not focused on algorithmic detection.
- Non-HRI focus or no reference to robots/agents.

- No use of DL methods (e.g., SVMs, random forests, traditional ML).
- Use of virtual reality/agents unrelated to HRI or engagement.
- Non-peer-reviewed works, book chapters, or editorials.
- Abstracts or extended abstracts (4 pages).
- Poorly written or lacking sufficient detail.
- Papers with no results or analysis.
- Redundant works presenting preliminary ideas followed by implementations.
- Low-quality publications or vague, incomplete studies.
- Scoping reviews, systematic reviews, or meta-analyses.

The focus of our study is on the application of DL methods for engagement detection in HRI. The rationale behind this decision was to concentrate on the advancements and specific contributions that Deep Learning approaches offer, particularly in terms of their ability to handle complex data representations and learn intricate patterns from large datasets. Even though the focus of our study was on Deep Learning methods, traditional methods have been compared to provide a better understanding of the overall landscape of engagement detection techniques. This comparison helps to contextualize the advancements made by DL algorithms and highlights the strengths and limitations of both approaches.

2.3. Screening

In the initial stage, papers that were not relevant to the topic of engagement detection or were off-topic were excluded from further consideration. After selecting initial papers based on the inclusion and exclusion criteria, in the second stage of screening, in the process of full-text reviewing, we identified papers that employed algorithmic development, including various DL algorithms, while separating traditional ML-based algorithms that are not DL (non-ANNs). Considering the inconsistent use of the term "Deep Learning" with respect to multi-layered neural networks, we used an inclusive interpretation of "Deep Learning" according to the use of Neural Networks. Within this broad scope, we included Neural Networks such as Autoencoders, Long Short-Term Memory Networks (LSTMs) and Gated Recurrent Units (GRUs), Convolutional networks (including Residual networks, Recurrent NNs (RNNs), Multilayer perceptrons, RCNN, YOLO, VGG-16/19, ResNet, Inception X), Deep Reinforcement Learning (including Deep Q Networks (DQN)), Bayesian Neural Networks and other algorithms that are based on Neural Networks. The selected papers were required to have sufficient details to allow for replication of the work, including algorithmic implementation or experimentation, and to have undergone basic empirical investigation. Abstracts/extended abstracts with no longer than four pages and papers that are poorly written or lack details were excluded from consideration.

To ensure the review was comprehensive, backward snowballing was implemented. This process involves examining the reference lists of selected papers to identify older, relevant studies. Data extraction was conducted by two extractors, with a third person serving as a conflict reviewer.

3. Results

In this section, we summarize the results of the conducted review in reference to the research questions outlined in the methodology. In total, 855 records were collected from the databases mentioned in 2.1, and after eliminating duplicates using the EPPI-Reviewer web screening platform (Thomas & Graziosi, 2010), 618 records remained for screening. EPPI Reviewer selects duplicate papers by using an automated algorithm to compare titles, authors, publication years, abstracts, and other bibliographic details to identify potential duplicates. Users can adjust the sensitivity of the matching criteria to balance between missing duplicates and incorrectly flagging unique records. The software groups potential duplicates into clusters based on similarities, which users then manually review to confirm or reject. This process ensures accurate and efficient identification and management of duplicates in systematic reviews. The selection process is depicted in Figure 1, adapted from the PRISMA-ScR flow diagram (Tricco et al., 2018).

Following the first screening stage, which was based solely on the title and abstract, 154 articles were included for further review. In the second stage, the full texts of the papers were reviewed and a total of 57 papers were identified as algorithmic development of engagement within which 36 papers employed DL approaches. 3 papers used non-text, 2 review papers, 7 were non-HRI related, 69 were non-algorithmic, 21 were non-ANNs (traditional algorithmic developments), 5 papers were without established results (more discussions of frameworks), and 11 Low quality and non-replicable papers were identified.

During the snowballing, an additional 37 papers were identified to undergo the selection process. Based on the title and abstract screening, 14 papers aligned with the predefined inclusion criteria. In the second stage of the review of these 14 records, 5 papers were identified as algorithmic development within which 2 papers employed DL algorithms, 8 were non-algorithmic papers, 3 were non-ANNs (traditional algorithmic developments), and also one review paper was identified.

In total, 38 papers were included in the final scope review list of papers that used DL methods for engagement detection in HRI (see Table 1). In the final included papers, there are some cases of multiple studies including many of the same authors, however, these works differ in one or more key aspects such as using a new dataset to train an engagement model, employing a different DL model, adopting an adaptive approach to interaction based on the developed engagement model, or utilizing a different method for data annotation. Out of these 38 papers that employed DL methods, 11 papers also employed traditional methods. Plus, 24 papers that were categorized as non-ANNs used traditional methods for engagement detection. In Figure 2, we can observe the papers that were extracted along with their respective publication year. The initial screening indicates

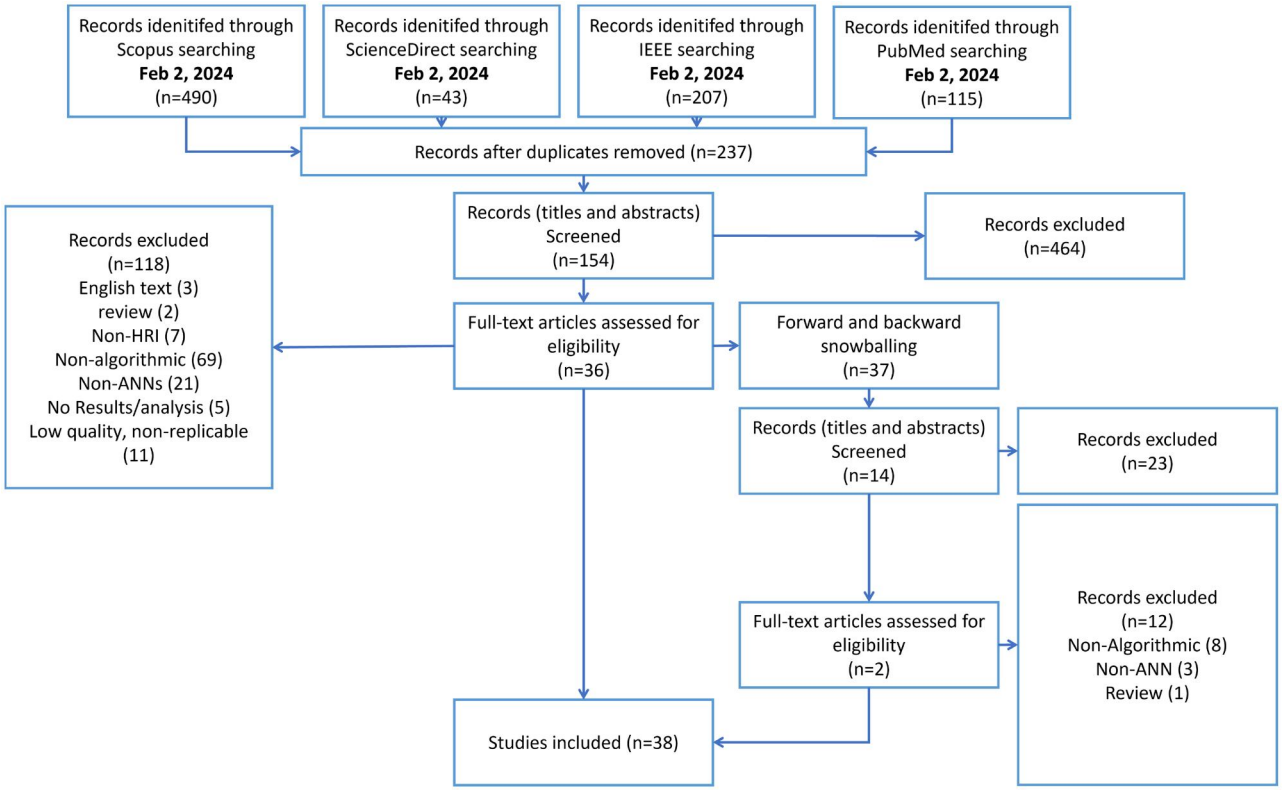

Figure 1. Screening process: been adapted based on the PRISMA-ScR flow diagram (Tricco et al., 2018).

the general trend of studies on engagement in HRI within which algorithmic-based publications for engagement detection are a sub-part. Notably, the analysis of algorithmic approaches for engagement detection reveals a rising preference for DL methods over the past five years compared to more traditional methods.

The review utilized a structured data extraction framework to systematically analyze studies on engagement detection in HRI. This framework included several key categories: the framework captured input features, detailing the specific sensory inputs used, such as facial features, gaze direction, and speech features, which are crucial for engagement detection. The algorithm type category identified the algorithms employed, distinguishing between various DL approaches. Furthermore, the framework differentiated between adaptation and non-adaptation approaches, categorizing models based on their application in adjusting to user engagement. The robot used in each study was specified to provide insights into practical applications, while the ground truth method of engagement documented how engagement was labeled, whether through expert annotation, self-reporting, or other techniques. By extracting data across these categories, the review aimed to create a comprehensive overview of methodologies and findings in the field, facilitating comparisons between studies and identifying trends, strengths, and gaps in current research.

3.1. Which DL methods are used for users' engagement modeling in HRI?

In the field of *engagement* research, various DL algorithms have been utilized for engagement detection, which can be categorized into CNN (ResNet3D, ResNetXt-50, ResNet, end-to-end 3D ConvNet, VGG19, ImageNet-VGG-F, EmoVGGFace2, VGGFace2-SA), RNN (BLSTM, GRU, RNN-CTC, ESNs, LSTM, LDN, ResNet-18), MLP (MLPR) and combinations of these types. As depicted in Figure 3, CNNs are the most frequently used method in the records (see Table 1 for more details).

CNNs stand out as the predominant modules for engagement detection, due to their ability to extract visual features from images. Visual cues, such as eye movements, body posture, facial features, and other non-verbal behaviors, play a crucial role in understanding and measuring user engagement. The preference for visual cues suggests that engagement is often perceived as a visually observable phenomenon rather than an internal state necessitating physiological sensory data. The focus on visual data may stem from the complexities involved in processing physiological signals or a lack of interest or expertise among researchers in that domain. Furthermore, most engagement studies concentrate on social engagement, where visible physical interaction is more significant than task engagement. In task engagement, participants are typically less physically expressive, making physiological or eye-tracking measures more valuable for assessing underlying engagement.

RNN-based models, given their inherent continuity and time-sensitive nature, particularly in contexts where temporal information is crucial for engagement modeling, especially emotional engagement that evolves over time (Scherer, 2005), are one of the other most prevalent used DL methods. For example, in scenarios requiring engagement detection using speech features, temporal information is essential (Pattar et al., 2019; Inoue et al., 2018; Rudovic et al., 2019).

Table 1. A summary of included publications, detailing authors, title and study type, input features, algorithm types, adaptability (adaptive/non-adaptive), robot or agent employed, and ground truth or dataset utilized.

Authors	Title/study type	Sensory inputs/features (facial expressions, gaze, speech, …)	Algorithm type	Adaptive/non-adaptive	Robot/agents	Ground truth (dataset)
(Mollahosseini et al., 2018)	Studying Effects of Incorporating Automated Affect Perception with Spoken Dialog in Social Robots **/Experimental**	Facial Features [expressions]	CNN [ResNet]	Adaptive [empathic conversational feedback, non-neural emotion expression]	**Ryan**	Dataset [AffectNet] Manual annotation [11 discrete emotion categories including basic emotions]
(Rossi & Rossi, 2021)	Engaged by a Bartender Robot: Recommendation and Personalization in Human-Robot Interaction	Facial Features [expressions, brow raise, brow furrowed, mouth open], Head Pose, Body Pose, Gaze	MLP [MLPR]	Non-Adaptive	**Pepper**	Data collection Manual annotation [by psychologists using the PARE – Pediatric Assessment Rehabilitation Engagement – Scale]
(Abdelrahman et al., 2022)	Multimodal Engagement Prediction in Multiperson Human–Robot Interaction	Facial Features, Gaze, Head Pose	CNN [ResNet-18, combined with rule-based approach]	Adaptive [Robot turning toward engaged participant, turning on the monitor, if disengagement detected a reminder timer starts]	**Cobot [robot arm]**	Data collection Manual annotation [according to events (users' instruction to robot, or interacting on touch screen]
(Bartlett et al., 2019)	What Can You See? Identifying Cues on Internal States From the Movements of Natural Social Interactions	Facial Features [landmarks], Head Pose, Body Pose	MLP	Non-Adaptive	**No robot**	Dataset [PInSoRo] Manual annotation [on-line study ratings by participants]
(Bartlett et al., 2021)	Estimating levels of engagement for social human-robot interaction using legendre memory units	Facial Features [landmarks], Body Pose [skeletal and hand landmarks]	RNN [LDN-MLP, LDN-MLP-KNN], MLP	Non-Adaptive	**No robot**	Dataset [PInSoRo], pre-labeled video for goal-oriented play', 'aimless play' and 'no play' as 'high', 'intermediate' and 'low' task engagement respectively
(Ben-Youssef et al., 2019	On-the-Fly Detection of User Engagement Decrease in Spontaneous Human–Robot Interaction Using Recurrent and Deep Neural Networks	User's Position [Distance], Gaze, Head Pose, Facial Features [expression], Speech Features.	MLP	Non-Adaptive	**Pepper**	Dataset [UE-HRI] Manual Annotation [by two annotators: a researcher who knew the purpose of the work and an uninformed one who did not]
(Dresvyanskiy et al., 2021)	Deep Learning Based Engagement Recognition in Highly Imbalanced Data	Facial Features [FAU]	CNN [EmoVGGFace2, VGGFace2-SA]-RNN[LSTM]	Non-Adaptive	**No robot**	Existing Dataset [DAiSEE dataset]
(Del Duchetto & Hanheide, 2022)	Are you still with me? Continuous Engagement Assessment from a Robot's Point of View	Facial Features, Head Pose, Body Pose, Users' Position	CNN [ResNetXt-50]-RNN[LSTM]	Adaptive [engagement as RL reward for improving social behavior]	**Lindsey**	Dataset [TOGURO dataset] Manual Annotation [3 annotators, inter-rater agreement]
(Atamna & Clavel, 2020)	HRI-RNN: A user-robot dynamics-oriented RNN for engagement decrease detection	User's Position [distance], Gaze, Head Pose, Facial Features, Speech Features, Robot's Speech Features	RNN[GRU]	Non-Adaptive	**Pepper**	Dataset [UE-HRI]
(Del Duchetto et al., 2020)	Learning on the Job: Long-Term Behavioural Adaptation in Human-Robot Interactions	Facial Features, Head Pose, Body Pose, Users' Position	CNN [ResNetXt-50]-RNN[LSTM]	Non-Adaptive	**Lindsey**	Dataset [TOGURO dataset] Manual annotation [Three expert annotators with Inter-Rater agreement]
(Dhamija & Boult, 2018)	Automated Action Units Vs. Expert Raters: Face off	Facial Features [FAU]	RNN	Non-Adaptive	**Pepper [dataset]**	Dataset [EASE, UE-HRI] Self-report
(Jain et al., 2020)	Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders **/Intervention**	Head Pose, Facial Features [landmarks, FAU], Gaze, Speech Features [frequency, intensity and harmonicity], Task Performance [challenge level, incorrect responses, elapsed time in a session]	MLP	Adaptive [RL based personalized feedback (speak) and instruction]	**Kiwi**	Data collection [7 children participants with ASD] Manual annotation [inter-rater reliability]

(*continued*)

Table 1. Continued.

Authors	Title/study type	Sensory inputs/features (facial expressions, gaze, speech, …)	Algorithm type	Adaptive/non-adaptive	Robot/agents	Ground truth (dataset)
(Poltorak & Drimus, 2017)	Human-robot interaction assessment using dynamic engagement profiles	Facial Features [expressions], Head Pose	CNN[ImageNet-VGG-F]	Non-Adaptive	**Unspecified robot**	Dataset [SFEW dataset, FER dataset] Data collection [4 participants, three female and one male, 50% aged 20–25 years old 50% 35–40 years old]
(Rajavenkatanarayanan et al., 2018)	Monitoring task engagement using facial expressions and body postures **/Experimental**	Facial Features [expression], Body Pose, Physiological Sensors [Muse sensor]	CNN[VGG19]	Non-Adaptive	**NAO**	Data collection [11 male participants (graduate students)] Tools [MUSE sensor]
(Saleh et al., 2021)	Improving users engagement detection using end-to-end spatio-temporal convolutional neural networks	Facial Features [FAU, landmarks], Head Pose	CNN [end-to-end 3D ConvNet], RNN [LSTM]	Non-Adaptive	**Pepper**	Dataset [UE-HRI]
(Duque-Domingo et al., 2020)	Gaze Control of a Robotic Head for Realistic Interaction With Humans	Speech Features, Gaze, Body Pose, User's Position	RNN[STM]	Adaptive [attention to user]	**Custom Robot Head**	Data collection [8 participants] Manual annotation
(Huang et al., 2017)	Speaker dependency analysis, audiovisual fusion cues and a multimodal BLSTM for conversational engagement recognition	Facial Features [EFI, HOG, Gabor Filter], Speech Features [pitch, MFCCs, Onset]	RNN[BLSTM]	Non-Adaptive	**No robot**	Existing dataset [Cardiff]
(Inoue et al., 2018)	Engagement recognition in spoken dialogue via neural network by aggregating different annotators' models	Speech Features [Backchannels, Laughing], Gaze, Head Pose [nodding]	RNN[GRU]	Non-Adaptive	**Erica**	Data collection [20 Japanese subjects (12 females and 8 males) from teenagers to over 70 years old] Manual annotation [inter-rater agreement]
(Silvia Ivani et al., 2022)	A gesture recognition algorithm in a robot therapy for ASD children	Body Pose [Skeleton]	CNN	Non-Adaptive	**NAO**	Data collection [22 participant, 9 healthy children and 13 adults of which 3 with ASD] [4 participants, ASD children for testing] Manual annotation
(Rudovic et al., 2019)	Multi-modal Active Learning From Human Data: A Deep Reinforcement Learning Approach **/Experimental**	Facial Features, Body Pose, Speech Features, Physiological Sensors [E4 wristband, galvanic skin conductance, body temperature, accelerometer data]	RNN [LSTM+RL-based active data labelling]	Non-Adaptive	**NAO**	Data collection [17/18 children, ages 3–13, with Japanese/European background, the cross-cultural dataset of children with ASC] Manual annotation [Experts]
(Suraj Prakash Pattar et al., 2019)	Intention and Engagement Recognition for Personalized Human-Robot Interaction, an integrated and Deep Learning approach	Facial Features [Expressions], Speech Features, Gaze, Head Pose, Body Pose [video]	RNN [within Autoencoder]	Non-Adaptive	**Pepper**	Data collection [8 participants, English speaking; Aged between 20 to 30; Able bodied; Persons studying or working on Robotics] Unsupervised
(Alghowinem et al., 2021)	Beyond the Words: Analysis and Detection of Self-Disclosure Behavior during Robot Positive Psychology Interaction **/Intervention**	Facial Features [landmarks] Gaze, Speech Features [yaw, pitch, roll, jitter, formants], Head Pose	MLP	Non-Adaptive	**Jibo**	Data collection [35 undergraduate students (age M = 18.94, SD = 1.43; 27 female, 7 male, and 1 other] Manual annotation [experts]
(Kawahara et al., 2018)	Audio-Visual Conversation Analysis By Smart Posterboard and Humanoid Robot	Facial Features, Gaze, Speech Features, Head Pose, Body Pose	RNN [LSTM]-CTC	Non-Adaptive	**Erica**	Data collection [20 subjects, 20 sessions of interacting with robot] Manual annotation
(T. Liu & Kappas, 2018)	Predicting Engagement Breakdown in HRI Using Thin-Slices of Facial Expressions	Facial Features [FAU]	RNN [Echo State Networks (ESNs)]	Non-Adaptive	**Pepper**	Dataset [UE-HRI]

(*continued*)

Table 1. Continued.

Authors	Title/study type	Sensory inputs/features (facial expressions, gaze, speech, …)	Algorithm type	Adaptive/non-adaptive	Robot/agents	Ground truth (dataset)
(Vaufreydaz et al., 2016)	Starting engagement detection towards a companion robot using multimodal features	Facial Features, Gaze, Speech Features [Laughter, vocalization], Head Pose, Body Pose, User's Position [distance]	MLP-SVM	Non-Adaptive	**Kompaï**	Data collection [19 participants were from 20 to 35 years old, almost 50% male/50% female, students, administrative assistants and researchers] Semi-Automatic [with a human expert confirming]
(Zhang et al., 2021)	Engagement Intention Estimation in Multiparty Human-Robot Interaction	Facial Features, Gaze, Head Pose, Body Pose, User's Position	CNN-RNN [LSTM]	Non-Adaptive	**Pepper**	Dataset [ATC Trajectory, JPL-Interaction, UE-HRI]
(Zhang et al., 2022)	Engagement estimation of the elderly from wild multiparty human–robot interaction	Facial Features, Gaze, Head Pose, Body Pose, User's Position	CNN [ResNet3D]-[Self-Attention] + [GAT]	Non-Adaptive	**Unspecified robot**	Dataset [BHEH] Manual annotation
(Sümer et al., 2023)	Multimodal Engagement Analysis From Facial Videos in the Classroom	Facial Features [facial expressions, action units], Head pose	MLP, RNN (LSTM)	Non-Adaptive	**No robot**	Dataset [Affect-Net, Attention-Net] Manual annotation [of collected Data, two raters]
(Bandi & Thomas, 2023)	A New Efficient Eye Gaze Tracker for Robotic Applications	Facial Features [full-face]	CNN	Non-Adaptive	**Pepper**	Dataset [PIIFaceGaze, ETH-XGaze]
(Shenoy et al., 2022)	A Self Learning System for Emotion Awareness and Adaptation in 2Humanoid Robots	Facial Features	CNN [ResNet50, Inception v3]	Adaptive	**NAO**	Dataset [NIMH-ChEFS, LIRIS-CSE, DEFSS, Dartmouth, AffectNet] Data collection [75 participants from the undergraduate and graduate student] Self-report
(Kokate et al., 2022)	An Algorithmic Approach to Audio Processing and Emotion Mapping	Speech Features [Mel-spectrogram, MFCCs, Chromogram]	CNN [1D.CNN,2D-CNN] Transformers, MLP	Non-adaptive	**No robot**	Dataset [RAVDESS, SAVEE, Urdu dataset] Self-report
(Simões et al., 2023)	Deep-Learning Based Classification of Engagement for Child-Robot Interaction	Facial Features [expressions]	CNN [ResNet18, VGG16, ResNet50, ConvNeXt]	Non-adaptive	**NAO**	Dataset [CAFÉ, FER2013] Data collection [3 children for testing purposes] Manual annotation
(Nezami et al., 2020)	Automatic Recognition of Student Engagement using Deep Learning and Facial Expression	Facial Features [expressions]	CNN [VGG-B]	Non-adaptive	**Virtual Environment [Omosa]**	Dataset [Engagement Recognition [ER], FER2013] Data collection 20 students [two public secondary schools involving twenty students 11 girls and 9 boys] Manual annotation [undergraduate Psychology students]
(Weng et al., 2022)	Development of a Visual Perception System on a Dual-Arm Mobile Robot for Human-Robot Interaction.	Facial Features [expressions], Speech Features	CNN + NLP Modules [google NLP]	Adaptive	**Mobi**	Dataset [FER2013]
(Dhaussy et al., 2023)	Interaction acceptance modelling and estimation for a proactive engagement in the context of human-robot interactions	Facial features, Head Pose, Body Pose	RNN [GRU], CNN	Non-adaptive	**Unspecified robot**	Data collection [12 individuals {all aged in (20, 30)}] Manual annotation [two annotators labeled on a scale of 1 to 5]
(Świetlicka et al., 2023)	Graph Neural Networks for Natural Language Processing in Human-Robot Interaction	Speech Features [turn taking, …]	CNN [GCN (Graph Convolutional Networks)]	Non-adaptive	**NAO**	Dataset [MHHRI] Self-Report [Big Five Personality Traits]

(*continued*)

Table 1. Continued.

Authors	Title/study type	Sensory inputs/features (facial expressions, gaze, speech, …)	Algorithm type	Adaptive/non-adaptive	Robot/agents	Ground truth (dataset)
(Gonzalez & Mizuuchi, 2023)	Physical Embodiment versus Novelty – which influences interactions with Embodied Conversational Agents more?	Facial Features [expressions]	CNN [VGG16, ResNet18]	Non-adaptive	**Social Plantroid**	Dataset [AffectNet] Data collection [27 participants[14 were female and 13 were male, mean age of 25.56 years (standard deviation: 4.05 years), (19 – 33 years old)]
(Rakhymbayeva et al., 2022)	To Transfer or Not To Transfer: Engagement Recognition within Robot-assisted Autism Therapy	Facial Features [landmarks, expressions], body pose,	CNN	Non-adaptive	**No robot**	Dataset [Qamqo, PInSoRo]

Nonetheless, due to inadequate datasets or the challenges associated with ground truth establishment of engagement, RNN-based models have not been extensively investigated, presenting a promising avenue for future research. Six papers used a combination of CNNs and RNNs to capture visual cues (Del Duchetto & Hanheide, 2022; Zhang et al., 2021). Nine papers referred to using MLP methods. However, due to their limited capabilities, MLP methods have been primarily used for low-level feature extraction (e.g., Jain et al., 2020; Kokate et al., 2022; Sümer et al., 2023; Rossi & Rossi, 2021) or in conjunction with other methods or pre-trained modules (e.g., Bartlett et al., 2019; Ben-Youssef et al., 2019). Additionally, one publication utilized transformers (Kokate et al., 2022) which show promise for future engagement detection due to their ability to capture temporal differences and integrate multimodal data—such as visual, auditory, and textual cues.

Papers utilizing traditional methods and not involving DL were separately categorized as non-ANNs. Papers utilizing both DL and traditional methods were classified as both DL and non-ANNs. Although traditional methods like SVM (Support Vector Machine), RF (Random Forest), LR (Logistic Regression), DT (Decision Trees), K-NN (K Nearest Neighbors), BN (Bayesian Networks), and HMM (Hidden Markov Model) remain popular for engagement detection, these methods often require dimension reduction pre-processing (Boccanfuso et al., 2016; Sümer et al., 2023) or feature extraction prior to the training process. For instance (Alyüz et al., 2016), leveraged the Intel RealSense SDK to extract features from video data for input into an RF algorithm. Additionally, others utilized the OpenCV packages to extract desired features (e.g., Haar-like features) for modeling purposes (Prado et al., 2011; Salam et al., 2017; Sanghvi et al., 2011). Furthermore, certain studies employed Kinect sensors to extract skeleton and body motion features (Foster et al., 2017; Tuyen et al., 2018). Although traditional methods are practical and straightforward for engagement detection, especially in scenarios with limited data, DL-based methods demonstrate efficiency in feature extraction, mitigating the need for specific feature assumptions, however, they require a substantial amount of well-annotated data.

3.2. Which sensory inputs are used for users' engagement modeling using DL methods in HRI?

Sensory inputs provide valuable information about how scholars interpret engagement within HRI. By analyzing diverse sensory data used for training DL techniques, researchers can gain a deeper understanding of how users engage with robots, in order to develop more personalized and adaptive interaction experiences. Different sensory inputs offer complementary cues that can help in accurately detecting and interpreting users' engagement states. By leveraging multiple modalities such as facial features, body poses, gaze patterns, and speech features, DL models can capture nuanced aspects of user engagement, leading to more robust and reliable engagement detection systems in HRI scenarios.

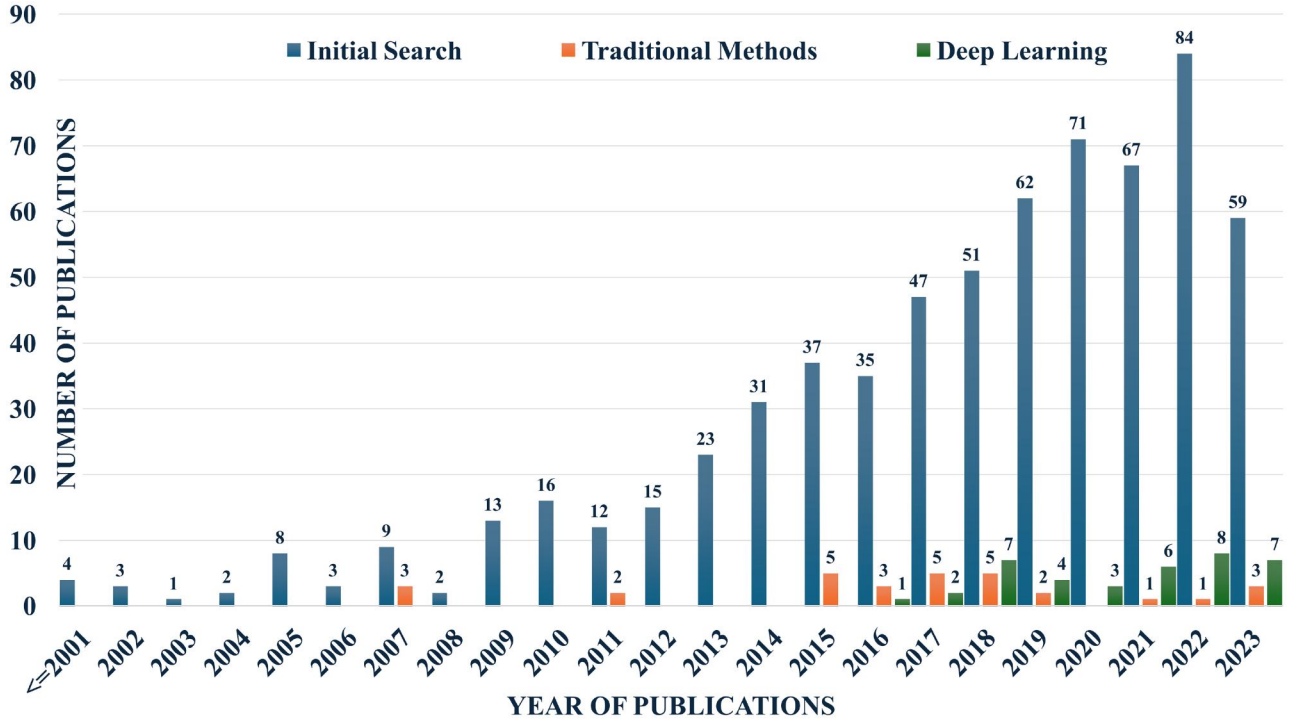

Figure 2. Number of publications per year: search and snowballing resulted publications: publications that used traditional methods [non-neural network such as SVM, Decision Trees, etc.] for engagement modeling and publications that used DL algorithms for modeling purposes.

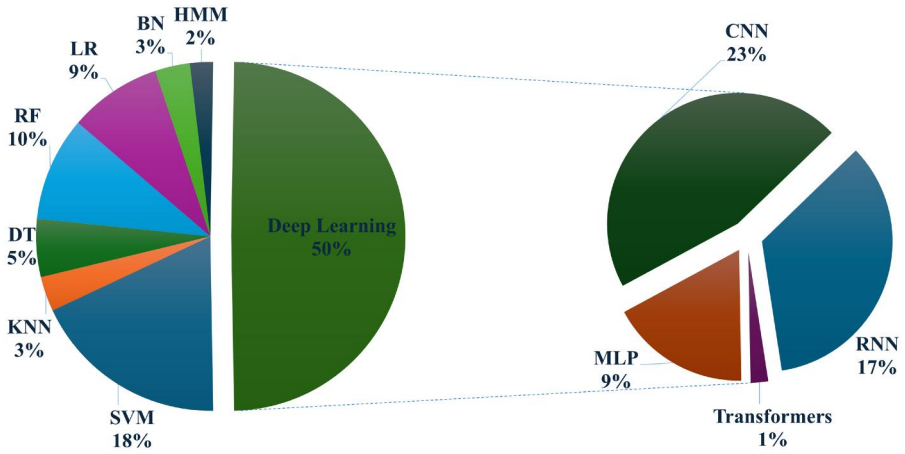

Figure 3. Types of algorithms used in the publications.

Figure 4 illustrates various features used by these publications for engagement detection, considering that a majority of studies employed multimodal features (see Table 1). These features encompass various features used in both traditional methods and DL methods including facial features (such as Facial Action Units, facial landmarks, Histogram of oriented gradients (HOG), and Gabor Filter), head pose (including nodding), body pose (involving skeletal, hand movement, or landmark, and Kinect data), gaze, speech features (including laughter, vocalization, yaw, pitch, roll, jitter, formants, backchannels, frequency, intensity, and harmonicity), user position (distance), physiological sensors (such as EEG tools like MUSE, E4 wristband, galvanic skin conductance, body temperature, accelerometer data, etc.), and task performance which includes challenge level, incorrect responses, and elapsed time in a session (see Table 1). Facial features (33 papers), head pose (19 papers), and body pose (16 papers) emerge as the most frequently employed features for engagement detection using DL methods (Figure 4). The use of facial features and gaze is notably lower in traditional methods compared to DL methods. This is directly related to the low potential of traditional methods for extracting high-level features like facial features unless they utilize pre-trained modules. Speech features are also utilized in 14 studies employing DL for engagement detection. Their usage is limited in comparison to facial features, head poses, and body poses. Additionally, eight studies incorporated user position as a feature for engagement detection using DL methods. However, user position offers limited insights for engagement detection as it mainly indicates whether engagement is established or not rather than measuring given information about engagement's quality or intensity, for example (Ben-Youssef

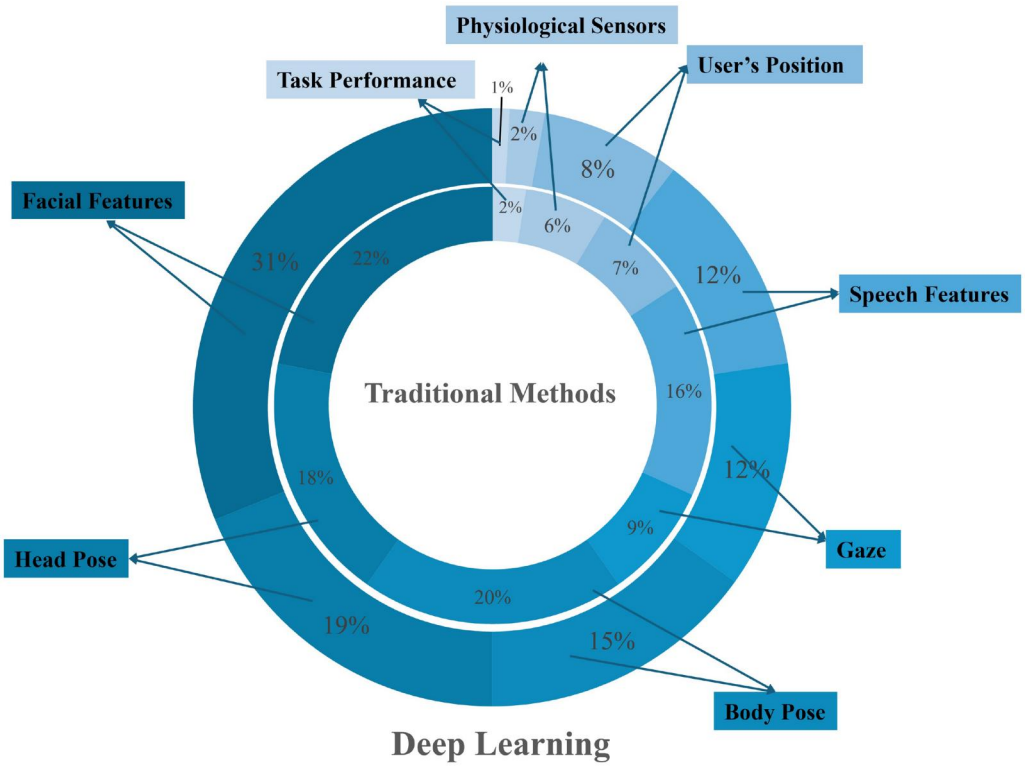

Figure 4. Key sensory inputs/features used for training traditional methods [non-neural network such as SVM, Decision Trees, etc.] and DL methods.

et al., 2019), used users' position (distance from robot) to establish engagement ground truth.

The utilization of physiological data for engagement detection using DL methods remains uncommon, likely due to the complexity of pre-processing and quantifying the often noisy sensory data. Only two papers used physiological data for engagement detection using DL methods (Rajavenkatanarayanan et al., 2018; Rudovic et al., 2019). Similarly, only one paper used task performance in order to detect engagement using DL methods (Jain et al., 2020). The limited usage of task performance or action-related data in engagement detection using DL methods suggests a gap in leveraging contextual information for engagement evaluation. This gap may be attributed to challenges in obtaining appropriate labels for different task performances in relation to other sensory data, which could explain the limited literature utilizing performance data. It is worth noting that it is more frequent to use traditional methods with physiological data. Five papers that utilized traditional methods for engagement detection did employ physiological sensory data (Fan et al., 2021; Liu et al., 2007; Mower et al., 2007; Ramadurai et al., 2024; Rani & Sarkar, 2007).

3.3. How does the literature address the establishment of ground truths for engagement in HRI for training DL models?

The choice of the engagement ground truth establishment method directly impacts the quality and reliability of the training data used to train DL models for engagement detection. By comprehensively examining the prevalent engagement ground truth establishment methods, researchers can make informed decisions about how to label their datasets, ensuring that the models are trained on accurate and representative data that captures the nuances of user engagement in HRI scenarios. This understanding is crucial for improving the performance and robustness of DL models in detecting and interpreting engagement signals, ultimately enhancing the overall user experience in HRI. Moreover, knowledge of common engagement ground truth establishment methods enables researchers to compare and benchmark their own annotation strategies against established practices in the field. By understanding the strengths and limitations of different annotation approaches, researchers can evaluate the effectiveness of their annotation processes and make informed decisions about how to optimize their data labeling procedures.

The primary method for labeling data in the records is the manual annotation by experts (17 out of 38 papers), who have specialized knowledge relevant to the established annotation process by researchers. Four studies have utilized users' self-reports as ground truth (Dhamija & Boult, 2018). Additionally, four studies have employed unsupervised methods to extract engagement patterns (Pattar et al., 2019). used clustering to classify data into four categories (approaching, interacting, leaving, and uninterested), with a timestamp serving as the basis for comparison (Rajavenkatanarayanan et al., 2018). used the EEG data collected from the MUSE sensor to measure the engagement level of the users during the task. This engagement value computed from the MUSE sensor data was considered the ground truth for the engagement detection model (Vaufreydaz et al., 2016). used semi-automatic labeling for engagement detection by initially segmenting events based on timestamped notes from the experimenter. Automatic

labeling was then applied using tablet interaction data and available features to assign labels such as "will interact" for approaching the robot and "leave interact" for paths taken after disengaging. All labels were reviewed by a human expert who examined video recordings to ensure accuracy and consistency in the labeling of the dataset. Abdelrahman et al. (2022) extracted the ground truth of users' engagement based on the tasks such as registration, interaction, and answering questionnaires. By tracking these engagement and disengagement actions during the study, the researchers were able to label the data events accordingly.

Twenty papers have leveraged datasets to develop engagement models, particularly evident in DL-based algorithms (see Figure 5). The availability of datasets has contributed to the growth of using DL-based models for engagement detection. While using datasets is the most common approach in DL publications, human experts' annotation is the most frequent ground-truth labeling in traditional methods. It's noteworthy that while these datasets may not always be labeled for engagement specifically, they often contain labels for affect detection which is a component of engagement as defined by Doherty and Doherty (2019). In total, thirty-five papers incorporated facial features for engagement detection which underscores the importance of the affect component of engagement.

3.4. What datasets have been used for training DL models in HRI?

By examining the datasets used in prior studies, researchers can gain insights into the types of data modalities, features, and annotations that have proven effective for training engagement detection models. This knowledge enables researchers to make informed decisions when selecting or creating datasets for their own investigations, ensuring that their models are built on a solid foundation of relevant and representative data. The diversity of engagement datasets used in HRI research highlights the range of scenarios and contexts in which engagement detection models have been developed and tested.

In total 12 datasets have been used to detect user engagement in HRI. These include UE-HRI dataset (used in five publications), PlnSoRo (used in three publications), FERdataset (used in three publications), and TOGURO (used in three publications). Other datasets such as BHEH, ATC Trajectory, JPL-Interaction dataset, Cardiff, SFEW dataset, EASE, DAiSEE dataset, and Affectnet dataset have been used in one publication each. While most of these datasets are publicly accessible, researchers often need to register or request access based on the dataset owner's policies.

The UE-HRI dataset was obtained from the interaction of the Pepper robot with people in a public exhibition. This dataset includes various features such as video, voice, sonar, and laser data, and has been manually annotated to include cues of engagement decrease and negative affect. The TOGURO dataset is collected from the interaction of the NAO robot in a public place and includes features such as video streams and cues of verbal/non-verbal behaviors of users, as well as information on users' positions. Additionally, several affect-based datasets have been utilized in engagement research, including SFEW, FER, and AffectNet (Mollahosseini et al., 2018; Poltorak & Drimus, 2017). Despite the existing datasets, there is a gap in well-structured datasets for engagement detection that combines different components of engagement including affect, and attention.

Although, leveraging a combination of these affect datasets with tailored labeled data from specific HRI setups could offer a promising approach to address engagement

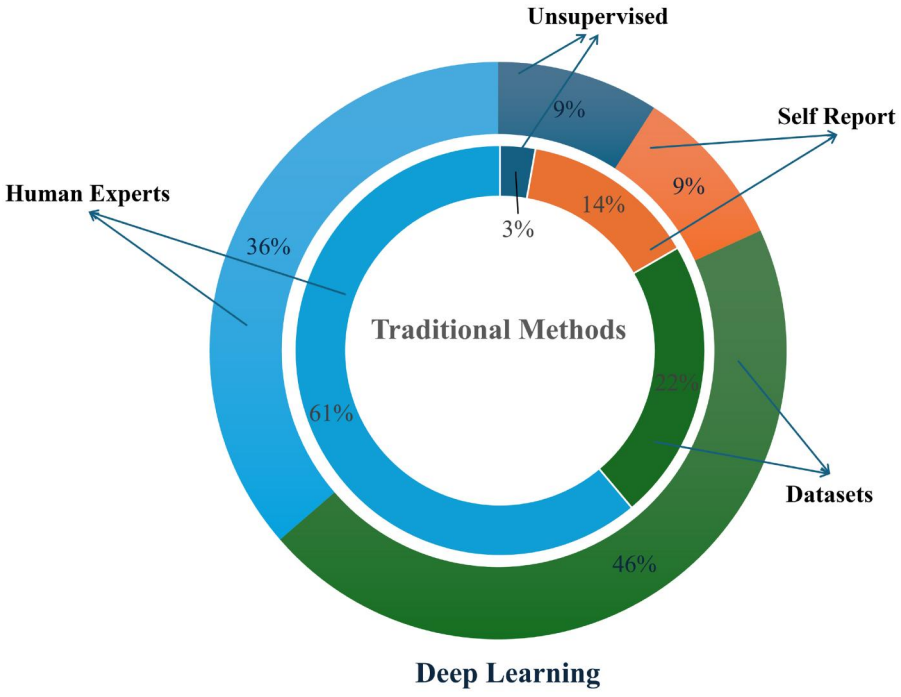

Figure 5. Labeling approaches for data preparation for engagement modeling.

detection, establishing suitable datasets for engagement analysis is time-intensive, and without clear definitions, developing universal datasets for all HRI setups is challenging. Different datasets might generalize over different use cases and how specific they are to the robot or interactive domain they are utilized on, for example, the UE-HRI dataset is developed using the Pepper robot in a public space where the annotation focused on a specific definition of engagement that is tied with the distance of users from the robot, user's head orientations and speech signals (Ben-Youssef et al., 2017). Another critical consideration in dataset development is the targeted demographic. Notably, studies like Jain et al. (2020) and Silvia Ivani et al. (2022) emphasize intervention-based data collection involving children with Autism Spectrum Disorder (ASD). For such groups, creating engagement detection models for social robots necessitates expert annotation of participants' engagement levels (Rakhymbayeva et al., 2022).

3.5. What types of HRI setups, such as social robots, virtual agents, or interactive environments, are used to develop engagement detection models?

Understanding the types of HRI setups used to develop engagement detection models provides insights into the diversity of platforms and technologies employed in HRI research, showcasing the versatility and applicability of engagement detection across different contexts. By identifying the specific setups utilized, researchers can gain a better understanding of the environments in which engagement detection models are tested and validated, leading to more informed decisions regarding the generalizability and scalability of these models. Furthermore, knowing the types of HRI setups used for engagement detection helps researchers and practitioners tailor their approaches and methodologies to suit specific interaction scenarios. Different setups may require unique considerations in terms of sensor modalities, data collection methods, and algorithmic approaches, highlighting the importance of adapting engagement detection techniques to the characteristics of the interaction environment.

The NAO robot is one of the most frequently used robots referred to in the extracted literature. Many of these studies have focused on the robot's physical movement as a means of interaction, with the NAO robot often being utilized to provide instructional feedback. For example Rajavenkatanarayanan et al. (2018), employed NAO robots to give instructions and provide verbal feedback based on user performance. Another commonly used robot is the Pepper robot, which is typically utilized in public spaces. Figure 6 depicts the robots that are being utilized in the extracted articles that utilized DL methods and traditional methods for engagement detection. It is worth noting that from the corpus of articles analyzed, it was observed that a subset of publications did not utilize any robots or datasets directly associated with robots and a few papers did not specify the type of robot they used (Bartlett et al., 2019; Dresvyanskiy et al., 2021; Huang et al., 2017; Poltorak & Drimus, 2017). Also, there are instances of publications that utilized engagement functions to control or adapt non-social robots. For example Pattar et al. (2019), used Cobots[Robotic arms] to interact with subjects and gathered data for engagement modeling (Mower et al., 2007; Rani & Sarkar, 2007). used two-wheel adaptive robots to interact with subjects and these robots were able to adapt to users' engagement levels, based on physiological sensory data. Moreover, two

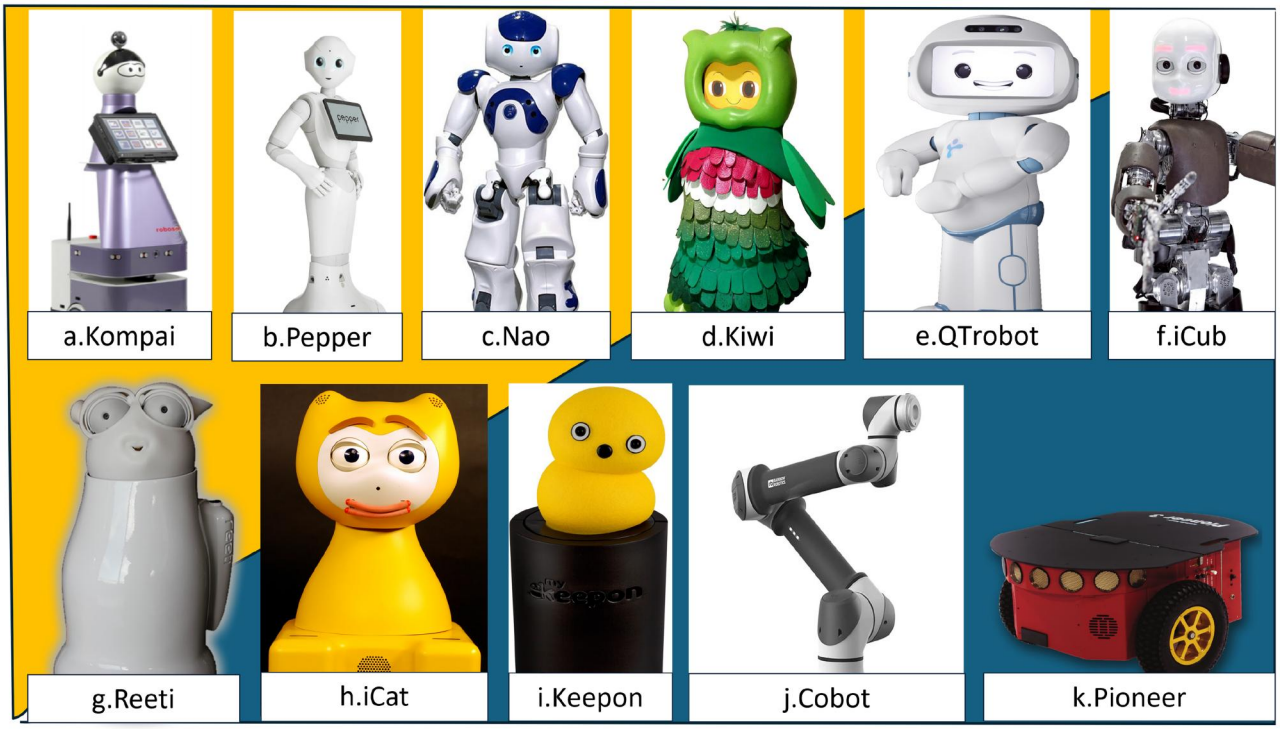

Figure 6. An array of robot designs featured in engagement-related publications, from humanoid and animal-like models to functional types like wheeled robots and robotic arms.

publications used users' interaction with virtual environments/agents to model engagement (Nezami et al., 2020; Trinh et al., 2018) using DL. Overall the result suggests a desire to enhance interaction by incorporating engagement detection, highlighting the significance of adjusting to users' engagement levels and a noticeable gap in the advancement of engagement methods with the aim for adaptive non-social robots and virtual agents, signaling an area that requires further investigation.

3.6. What engagement adaptation methods have been applied to the HRI setup?

Engagement adaptation in HRI plays a pivotal role in shaping the quality of interactions between humans and robots. By dynamically adjusting robots' behaviors in response to users' engagement levels, robots can create more personalized and engaging experiences that cater to individual preferences and communication styles. This adaptability not only enhances user satisfaction and enjoyment but also fosters effective communication and mutual understanding between humans and robots. Additionally, sustaining long-term engagement through continuous adaptation to users' engagement signals is crucial for maintaining users' interest and involvement over time. By dynamically adjusting their behaviors to match users' changing engagement levels, robots can ensure ongoing interaction and collaboration, preventing disengagement and promoting sustained user participation. This long-term engagement fosters a sense of continuity and connection in HRI, contributing to a positive and fulfilling user experience that aligns with users' evolving needs and preferences.

Seven out of the thirty-eight included publications employed adaptive approaches for interaction. Five of these papers utilized rule-based decisions that vary based on user engagement levels (Abdelrahman et al., 2022; Duque-Domingo et al., 2020; Mollahosseini et al., 2018). Two studies incorporated engagement levels into reinforcement learning (RL) based policy learners, which enables more tailored adaptations for users, widening the spectrum of adaptivity (Del Duchetto & Hanheide, 2022). employed RL policies to improve the social behavior of robots towards exhibition visitors, while (Jain et al., 2020) utilized RL-based feedback and instructions for adaptive behavior. The relatively low number of studies using adaptive models, especially those employing learning-based adaptations, indicates a potential research gap. This gap requires further exploration to enable more personalized interactions and improve overall HRI experiences.

4. Discussion

This scoping review serves as a comprehensive resource for researchers seeking to delve into the realm of DL techniques for engagement detection in HRI. By highlighting key findings, research gaps, and methodological considerations, the review sets the stage for future investigations aimed at enhancing the understanding and implementation of engagement detection mechanisms. The increasing adoption of DL algorithms for engagement detection signifies a shift towards data-driven approaches in understanding the engagement concept in HRI. By leveraging the capabilities of DL models researchers are able to extract meaningful insights from various data modalities like speech features, facial features, and physiological sensors to enhance engagement detection accuracy. The review emphasizes the importance of selecting appropriate sensory inputs/features and integrating different data modalities effectively to improve engagement detection using DL methods. Furthermore, the identification of common DL algorithms used for engagement modeling, such as CNNs like ResNet and VGG19, RNNs like BLSTM and LSTM, and MLPs, provides a roadmap for researchers to explore and compare different algorithmic approaches in their studies.

The scoping review shows that the detection of engagement using DL methods is often facilitated by the use of facial features and head and body pose as primary features (as demonstrated in Figure 4), while, speech and gaze features are also commonly utilized features. The selection of features employed is dependent on the specific definition of engagement being studied. There appears to be a lack of thorough exploration into using task performance features, physiological features, and user position in comparison to gaze, facial, and speech features. The scarcity of using physiological sensors is partly because of the delay usually associated with these sensory data which makes it impractical for instant engagement detection. Moreover, the emphasis on social engagement rather than task engagement contributes to this scarcity. While social engagement measurement relies more on visual signals, task engagement—especially in activities requiring minimal movement and where visual signals offer limited information—could benefit more from physiological signals despite the challenges in evaluating them.

It is noticeable that using visual cues for engagement modeling suggests a tendency to focus more on the visual aspect of engagement rather than the internal cues of engagement. Convolutional Neural Networks (CNNs) which dominate as the preferred algorithmic method for engagement detection facilitate this typical engagement interpretation. At the same time, there are studies that used Recurrent Neural Networks (RNNs) to capture temporal aspects of engagement. However, there's a need for more established methodologies for utilizing RNNs, particularly with datasets that consider temporal engagement aspects.

While traditional methods like Support Vector Machines (SVMs) remain favorable due to their simplicity in capturing relevant patterns, these methods often aided by feature selection techniques like PCA (Boccanfuso et al., 2016; Sümer et al., 2023) or leverage open-source tools such as OpenFace and OpenPose for facial and body cues extraction, driven by the scarcity of sufficient training data for feature extraction (Rudovic et al., 2019; Dhamija & Boult, 2018; Saleh et al., 2021). These methods are especially effective in scenarios where limited training data poses challenges for feature extraction in DL models. Furthermore, achieving real-time performance is critical in HRI, as delays in response can

diminish user experience; however, many DL models are computationally intensive, complicating this requirement.

Furthermore, annotating engagement is a challenging task that typically demands a significant time investment and at least two human annotators. Moreover, the lack of a clear and consistent definition of engagement further compounds the difficulty and necessitates training annotators to ensure consistent labeling making it even less appealing to develop engagement datasets. Unsupervised engagement labeling (Pattar et al., 2019) and semi-automatic annotation (Vaufreydaz et al., 2016) are potential methods to aid in the creation of engagement datasets. Incorporating both self-reports and expert annotators could also be a viable approach to address this issue (Rudovic et al., 2019). adopted an active data labeling strategy that omits unnecessary annotations by actively selecting frames for labeling based on an active learning approach. By tailoring the selection of data for each user, their engagement estimation model could be personalized to better suit the target child. The review recognizes that variability in ground truth establishment methods—such as manual annotation, self-reporting, and unsupervised techniques—could lead to inconsistencies in how engagement is defined and measured. Experts' annotation, while generally reliable, can vary based on the annotators' interpretations, potentially affecting the consistency of engagement labels. Similarly, self-reporting can introduce biases that may not accurately reflect actual engagement levels, leading to discrepancies across studies. The use of unsupervised methods, although beneficial for discovering patterns, may lack the nuanced understanding that human annotators provide, resulting in less reliable ground truth. By highlighting these variabilities, the review underscores the importance of establishing standardized methods for ground truth labeling to enhance the validity and comparability of results across different studies in the field of HRI.

Although various datasets have been employed to study engagement in HRI, they are often designed for specific applications and are not easily adaptable to other contexts. As mentioned, there are a number of papers that utilize the UE-HRI dataset for the detection of engagement. However, there are papers that have used affect-based datasets that are not established for engagement detection purposes, which indicates the inadequacy of appropriate datasets for this purpose. Notably, studies employing DL methods tend to rely more on datasets and this is mainly because of the substantial training requirements of DL approaches.

It is noteworthy that the robots featured in these studies primarily belong to the category of social/humanoid robots. This emphasis is driven by the potential for improving HRI through social interaction using engagement detection modules. However, while a few studies include non-social robots, there remains a gap in the literature regarding engagement modeling in non-social robots. Addressing this gap could significantly improve interactions with these robots.

Moreover, by recognizing the distinctions between public and targeted interactions, researchers can address diverse aspects of engagement, such as social behaviors, communication dynamics, and user preferences. One illustration of public interactions with robots involves the use of the Pepper robot in a public space to interact with random attendees (Ben-Youssef et al., 2017). The study collected the UE-HRI dataset which is widely utilized for investigating engagement (Zhang et al., 2021; Dhamija & Boult, 2018; Saleh et al., 2021; Liu & Kappas, 2018; Atamna & Clavel, 2020; Del Duchetto et al., 2020). Another approach relates to intervention or participation-based interactions, in which robots interact with individuals in a more targeted and time-specific manner (Alghowinem et al., 2021; Jain et al., 2020). The NAO humanoid robot is the most frequently used robot in this category (Anagnostopoulou et al., 2021). In the first approach, features such as individuals approaching or turning toward the robot, body posture, and gaze toward the robot, are the main indicators of engagement, while in the second approach, the focus is on the quality of interaction between the subject and the robot.

This review highlights some challenges and limitations in the existing literature. Despite the progress made in utilizing DL for engagement detection, challenges such as annotation complexities, dataset availability, and model generalizability remain prevalent in the field. Addressing these challenges through standardized methodologies, benchmark datasets, and cross-validation techniques will be crucial for advancing the reliability and applicability of DL-based engagement detection systems in real-world HRI scenarios. Future research efforts should focus on addressing these challenges to advance the field and facilitate the development of more effective and reliable engagement detection systems. Furthermore, conducting longitudinal studies to assess the long-term performance of engagement detection, can contribute to the development of more robust and adaptive engagement detection algorithms, ultimately enhancing the user experience and fostering deeper connections between humans and robots.

The effectiveness of engagement detection models can vary significantly depending on the environment in which the robots operate. For instance, robots used in public spaces, such as the Pepper robot interacting with random attendees at exhibitions, must be designed to handle spontaneous and diverse interactions. These settings often involve a wide range of user demographics and engagement levels, requiring models that can adapt to varying social cues and behaviors. In contrast, robots employed in clinical settings, such as those assisting patients with specific needs, operate within a more structured and targeted framework. Here, the interactions are often more focused, with defined goals related to therapy or rehabilitation. The engagement detection models in these contexts must account for the unique dynamics of patient interactions, including emotional states and specific communication needs. By recognizing these differences, researchers can develop tailored datasets and methodologies that enhance the applicability and reliability of engagement detection systems across various HRI scenarios. This contextual relevance is essential for ensuring that engagement detection models are not only effective but also sensitive to the nuances of different interaction environments, ultimately leading to improved human-robot interactions.

Ethical considerations also play a crucial role in engagement model development, particularly concerning user privacy and the potential for biases in the models can lead to inequitable treatment of different user demographics. Furthermore, there is a risk of over-reliance on AI systems in sensitive areas like healthcare, where the human element is vital for effective care. Despite these challenges, the accurate detection of user engagement through DL can lead to significantly improved user experiences, as SARs can tailor their interactions based on real-time engagement cues. This adaptability not only fosters more personalized and meaningful interactions but also enhances user satisfaction and trust, ultimately promoting the long-term adoption of SARs in various settings, including therapeutic and assistive environments. By addressing these challenges, researchers can pave the way for more effective and reliable engagement detection systems that enhance the overall impact of SARs on users' lives.

5. Conclusion

The primary objective of this scoping review paper is to provide a comprehensive overview of the current DL methods used for detecting user engagement in HRI, as well as the features employed for this purpose. The utilization of DL algorithms for engagement *detection* is a noticeable and continuously increasing trend in comparison to traditional methods. This trend is expected to gain further momentum in the near future with the development of more appropriate datasets. Through the use of a rigorous methodology, this review paper identified common features and patterns obtained by DL techniques for user engagement detection. The review reveals that DL techniques employ facial features and head poses as prominent features in detecting engagement. The most commonly utilized robots in this research area are social robots. The review also reported prevalent datasets and labeling approaches for developing engagement detection models.

It's important to note that while other related studies may have explored similar topics without explicitly addressing "engagement," this paper emphasizes algorithmic approaches for engagement detection. We identified gaps in the literature, particularly the need for more exploration into the temporal nature of engagement, the lack of context-based datasets, and a notable gap in investigating engagement for non-social robots. Furthermore, while there are studies that have utilized physiological sensory and performance-related data, there has been an insufficient investigation into employing these data for engagement detection and deriving substantive insights into the potential of this approach.

Authors' contributions

Conceptualisation: BSR. Methodology: BSR, IK, PG, RL. Formal analysis: BSR, IK, PG, RL. Investigation: BSR, IK, PG, RL. Data curation: BSR, IK, PG, RL. Writing—original draft: BSR. Writing—review and editing: BSR, IK, PG, RL. Project administration: BSR.

Disclosure statement

The authors declare that there are no conflicts of interest.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Bahram Salamat Ravandi http://orcid.org/0009-0008-2712-8684
Imran Khan http://orcid.org/0000-0003-4908-5001
Pierre Gander http://orcid.org/0000-0002-0214-7511
Robert Lowe http://orcid.org/0000-0002-0307-3171

References

Abdelrahman, A. A., Strazdas, D., Khalifa, A., Hintz, J., Hempel, T., & Al-Hamadi, A. (2022). Multimodal engagement prediction in multi-person human–robot interaction. *IEEE Access*, *10*, 61980–61991. https://doi.org/10.1109/ACCESS.2022.3182469

Ahmad, M. I., Mubin, O., Shahid, S., & Orlando, J. (2019). Robot's adaptive emotional feedback sustains children's social engagement and promotes their vocabulary learning: A long-term child–robot interaction study. *Adaptive Behavior*, *27*(4), 243–266. https://doi.org/10.1177/1059712319844182

Alghowinem, S., Jeong, S., Arias, K., Picard, R., Breazeal, C., & Won Park, H. (2021). *Beyond the words: Analysis and detection of self-disclosure behavior during robot positive psychology interaction* [Paper presentation]. 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 01–08). https://doi.org/10.1109/FG52635.2021.9666969

Alyüz, N., Okur, E., Oktay, E., Genc, U., Aslan, S., Mete, S. E., Stanhill, D., Arnrich, B., & Esme, A. A. (2016). Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1: 1 learning scenario? In *UMAP 2016 Extended Proceedings: Late-breaking Results, Posters, Demos, Doctoral Consortium and Workshops* (Vol. 1618). CEUR-WS.org. http://ceur-ws.org/Vol-1618/

Amaro, B., Silva, V., Soares, F., & Sena Esteves, J. (2019). Building a behaviour architecture: An approach for promoting human-robot interaction. In J. Machado, F. Soares, & G. Veiga (Eds.), *Innovation, engineering and entrepreneurship* (pp. 39–45). Springer International Publishing.

Amirova, A., Rakhymbayeva, N., Yadollahi, E., Sandygulova, A., & Johal, W. (2021). 10 years of human-nao interaction research: A scoping review. *Frontiers in Robotics and AI*, *8*, 744526. https://doi.org/10.3389/frobt.2021.744526

Anagnostopoulou, D., Efthymiou, N., Papailiou, C., & Maragos, P. (2021). *Engagement estimation during child robot interaction using deep convolutional networks focusing on asd children* [Paper presentation]. 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3641–3647). https://doi.org/10.1109/ICRA48506.2021.9561687

Andriella, A., Torras, C., & Alenyà, G. (2020). Cognitive system framework for brain-training exercise based on human-robot interaction. *Cognitive Computation*, *12*(4), 793–810. https://doi.org/10.1007/s12559-019-09696-2

Atamna, A., & Clavel, C. (2020). *Hri-rnn: A user-robot dynamics-oriented rnn for engagement decrease detection* [Paper presentation]. Interspeech 2020 (pp. 4198–4202). https://doi.org/10.21437/Interspeech.2020-1261

Bandi, C., & Thomas, U. (2023). *A new efficient eye gaze tracker for robotic applications* [Paper presentation]. 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6153–6159). https://doi.org/10.1109/ICRA48891.2023.10161347

Bärenholdt, M., Jensen, L. C., Pedersen, J. E., & Petersen, E. (2020). How do situation awareness affect people's physical engagement with a robot?. In Companion of the 2020 ACM/IEEE International

Conference on Human-Robot Interaction, HRI '20 (pp. 116–118). Association for Computing Machinery.

Bartlett, M. E., Stewart, T. C., & Thill, S. (2021). *Estimating levels of engagement for social human-robot interaction using legendre memory units* [Paper presentation]. Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, HRI '21 Companion, New York, NY, USA (pp. 362–366). Association for Computing Machinery. https://doi.org/10.1145/3434074.3447193

Bartlett, M. E., Edmunds, C. E., Belpaeme, T., Thill, S., & Lemaignan, S. (2019). What can you see? Identifying cues on internal states from the movements of natural social interactions. *Frontiers in Robotics and AI*, *6*, 49. https://doi.org/10.3389/frobt.2019.00049

Belletier, C., Normand, A., & Huguet, P. (2019). Social-facilitation-and-impairment effects: From motivation to cognition and the social brain. *Current Directions in Psychological Science*, *28*(3), 260–265. https://doi.org/10.1177/0963721419829699

Ben-Youssef, A., Clavel, C., Essid, S., Bilac, M., Chamoux, M., & Lim, A. (2017). *Ue-hri: A new dataset for the study of user engagement in spontaneous human-robot interactions* [Paper presentation]. Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, New York, NY, USA (pp. 464–472). ACM. https://doi.org/10.1145/3136755.3136814

Ben-Youssef, A., Varni, G., Essid, S., & Clavel, C. (2019). On-the-fly detection of user engagement decrease in spontaneous human-robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics*, *11*(5), 815–828. https://doi.org/10.1007/s12369-019-00591-2

Björling, E. A., Rose, E., & Ren, R. (2018). Teen-robot interaction: A pilot study of engagement with a low-fidelity prototype. In Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18, New York, NY, USA (pp. 69–70). Association for Computing Machinery.

Boccanfuso, L., Wang, Q., Leite, I., Li, B., Torres, C., Chen, L., Salomons, N., Foster, C., Barney, E., Amy Ahn, Y., Scassellati, B., & Shic, F. (2016). *A thermal emotion classifier for improved human-robot interaction* [Paper presentation]. 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 718–723). https://doi.org/10.1109/ROMAN.2016.7745198

Chan, J., & Nejat, G. (2010). *Promoting engagement in cognitively stimulating activities using an intelligent socially assistive robot* [Paper presentation]. 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (pp. 533–538).

Del Duchetto, F., & Hanheide, M. (2022). Learning on the job: Long-term behavioural adaptation in human-robot interactions. *IEEE Robotics and Automation Letters*, *7*(3), 6934–6941. https://doi.org/10.1109/LRA.2022.3178807

Del Duchetto, F., Baxter, P., & Hanheide, M. (2020). Are you still with me? continuous engagement assessment from a robot's point of view. *Frontiers in Robotics and AI*, *7*, 116. https://doi.org/10.3389/frobt.2020.00116

Dhamija, S., & Boult, T. E. (2018). *Automated action units vs. expert raters: Face off* [Paper presentation]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 259–268). https://doi.org/10.1109/WACV.2018.00035

Dhaussy, T., Jabaian, B., & Lefevre, F. (2023). *Interaction acceptance modelling and estimation for a proactive engagement in the context of human-robot interactions* [Paper presentation]. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Los Alamitos, CA, USA (pp. 3061–3066). IEEE Computer Society. https://doi.org/10.1109/ICCVW60793.2023.00330

Doherty, K., & Doherty, G. (2019). Engagement in hci: Conception, theory and measurement. *ACM Computing Surveys*, *51*(5), 1–39. https://doi.org/10.1145/3234149

Dresvyanskiy, D., Minker, W., & Karpov, A. (2021). Deep learning based engagement recognition in highly imbalanced data. In A. Karpov & R. Potapova (Eds.), *Speech and computer* (pp. 166–178). Springer International Publishing.

Duque-Domingo, J., Gómez-García-Bermejo, J., & Zalama, E. (2020). Gaze control of a robotic head for realistic interaction with humans. *Frontiers in Neurorobotics*, *14*, 34. https://doi.org/10.3389/fnbot.2020.00034

Fan, J., Ullal, A., Beuscher, L., Mion, L. C., Newhouse, P., & Sarkar, N. (2021). Field testing of ro-tri, a robot-mediated triadic interaction for older adults. *International Journal of Social Robotics*, *13*(7), 1711–1727. https://doi.org/10.1007/s12369-021-00760-2

Feil-Seifer, D., & Mataric, M. J. (2005). *Defining socially assistive robotics* [Paper presentation]. 9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005 (465–468). https://doi.org/10.1109/ICORR.2005.1501143

Foster, M. E., Gaschler, A., & Giuliani, M. (2017). Automatically classifying user engagement for dynamic multi-party human–robot interaction. *International Journal of Social Robotics*, *9*(5), 659–674. https://doi.org/10.1007/s12369-017-0414-y

Gonzalez, A. G. C., & Mizuuchi, I. (2023). *Physical embodiment versus novelty - which influences interactions with embodied conversational agents more?* [Paper presentation]. 32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28-31, 2023 (pp. 675–682). IEEE. https://doi.org/10.1109/RO-MAN57019.2023.10309561

Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep learning. MIT Press. http://www.deeplearningbook.org.

Hadfield, J., Chalvatzaki, G., Koutras, P., Khamassi, M., Tzafestas, C. S., & Maragos, P. (2019). *A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task* [Paper presentation]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1251–1256). https://doi.org/10.1109/IROS40897.2019.8968443

Huang, Y., Gilmartin, E., & Campbell, N. (2017). *Speaker dependency analysis, audiovisual fusion cues and a multimodal blstm for conversational engagement recognition* [Paper presentation]. Interspeech 2017 (pp. 3359–3363). https://doi.org/10.21437/Interspeech.2017-1496

Inoue, K., Lala, D., Takanashi, K., & Kawahara, T. (2018). *Engagement recognition in spoken dialogue via neural network by aggregating different annotators' models* [Paper presentation]. Interspeech 2018 (pp. 616–620). https://doi.org/10.21437/Interspeech.2018-2067

Jain, S., Thiagarajan, B., Shi, Z., Clabaugh, C., & Matarić, M. J. (2020). Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics*, *5*(39), eaaz3791. https://doi.org/10.1126/scirobotics.aaz3791

Kawahara, T., Inoue, K., Lala, D., & Takanashi, K. (2018). Audio-visual conversation analysis by smart posterboard and humanoid robot. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6573–6577).

Kokate, V., Karmakar, A., Kapasi, M., & Chavan, H. (2022). *An algorithmic approach to audio processing and emotion mapping* [Paper presentation]. 2022 5th International Conference on Advances in Science and Technology (ICAST) (pp. 225–230). https://doi.org/10.1109/ICAST55766.2022.10039580

Kubota, A., Cruz-Sandoval, D., Kim, S., Twamley, E. W., & Riek, L. D. (2022). *Cognitively assistive robots at home: Hri design patterns for translational science* [Paper presentation]. 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 53–62). https://doi.org/10.1109/HRI53351.2022.9889442

Liu, C., Conn, K., Sarkar, N., & Stone, W. (2007). *Affect recognition in robot assisted rehabilitation of children with autism spectrum disorder* [Paper presentation]. Proceedings 2007 IEEE International Conference on Robotics and Automation (pp. 1755–1760). https://doi.org/10.1109/ROBOT.2007.363576

Liu, T., & Kappas, A. (2018). *Predicting engagement breakdown in HRI using thin-slices of facial expressions* [Paper presentation]. The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018, Volume WS-18 of AAAI Technical Report (pp. 37–43). AAAI Press.

Logan, D. E., Breazeal, C., Goodwin, M. S., Jeong, S., O'Connell, B., Smith-Freedman, D., Heathers, J., & Weinstock, P. (2019). Social robots for hospitalized children. *Pediatrics*, *144*(1), e20181511. https://doi.org/10.1542/peds.2018-1511

Luria, M., Reig, S., Tan, X. Z., Steinfeld, A., Forlizzi, J., & Zimmerman, J. (2019). Re-embodiment and co-embodiment: Exploration of social

presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference, DIS '19*, New York, NY, USA (pp. 633–644). Association for Computing Machinery.

Lytridis, C., Bazinas, C., Kaburlasos, V. G., Vassileva-Aleksandrova, V., Youssfi, M., Mestari, M., Ferelis, V., & Jaki, A. (2019). *Social robots as cyber-physical actors in entertainment and education* [Paper presentation]. *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)* (pp. 1–6). https://doi.org/10.23919/SOFTCOM.2019.8903630

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://hrcak.srce.hr/89395

Mollahosseini, A., Abdollahi, H., & Mahoor, M. H. (2018). *Studying effects of incorporating automated affect perception with spoken dialog in social robots* [Paper presentation]. 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 783–789). https://doi.org/10.1109/ROMAN.2018.8525777

Mower, E., Feil-Seifer, D. J., Mataric, M. J., & Narayanan, S. (2007). *Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements* [Paper presentation]. RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication (pp. 1125–1130). https://doi.org/10.1109/ROMAN.2007.4415249

Mucchiani, C., Cacchione, P., Johnson, M., Mead, R., & Yim, M. (2021). *Deployment of a socially assistive robot for assessment of covid-19 symptoms and exposure at an elder care setting* [Paper presentation]. 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) (pp. 1189–1195). https://doi.org/10.1109/RO-MAN50785.2021.9515551

Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2020). Automatic recognition of student engagement using deep learning and facial expression. In U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M Maathuis, & C. Robardet (Eds.), *Machine learning and knowledge discovery in databases* (pp. 273–289). Springer International Publishing. https://doi.org/10.1007/978-3-030-46133-1_17

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. https://doi.org/10.1002/asi.20801

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., & Peters, C. (2020). Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*, *7*. https://doi.org/10.3389/frobt.2020.00092

Pattar, S. P., Coronado, E., Ardila, L. R., & Venture, G. (2019). *Intention and engagement recognition for personalized human-robot interaction, an integrated and deep learning approach* [Paper presentation]. 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM) (pp. 93–98). https://doi.org/10.1109/ICARM.2019.8834226

Poltorak, N., & Drimus, A. (2017). *Human-robot interaction assessment using dynamic engagement profiles* [Paper presentation]. 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids) (pp. 649–654). https://doi.org/10.1109/HUMANOIDS.2017.8246941

Prado, J. A., Seneviratne, L. D., & Dias, J. M. M. (November 7-9, 2011). *Synthesis of emotions on a human-robot-interactive platform* [Paper presentation]. Proceedings of the IASTED International Conference on Robotics (Robo 2011), Pittsburgh, USA.

Rajavenkatanarayanan, A., Babu, A. R., Tsiakas, K., & Makedon, F. (2018). *Monitoring task engagement using facial expressions and body postures* [Paper presentation]. Proceedings of the 3rd International Workshop on Interactive and Spatial Computing, IWISC '18, New York, NY, USA (pp. 103–108). Association for Computing Machinery. https://doi.org/10.1145/3191801.3191816

Rakhymbayeva, N., Balgabekova, Z., Nurmukhamed, M., Burunchina, K., Johal, W., & Sandygulova, A. (2022). *To transfer or not to transfer: Engagement recognition within robot-assisted autism therapy* [Paper presentation]. 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 1002–1006). https://doi.org/10.1109/HRI53351.2022.9889577

Ramadurai, S., Gutierrez, C., Jeong, H., & Kim, M. (December, 2024). Physiological indicators of fluency and engagement during sequential and simultaneous modes of Human-Robot collaboration. *IISE Transactions on Occupational Ergonomics and Human Factors*, *12*(1–2), 97–111. https://doi.org/10.1080/24725838.2023.2287015

Rani, P., & Sarkar, N. (2007). Operator engagement detection for robot behavior adaptation. *International Journal of Advanced Robotic Systems*, *4*(1), 1. https://doi.org/10.5772/5716

Ritschel, H., Baur, T., & André, E. (2017). *Adapting a robot's linguistic style based on socially-aware reinforcement learning* [Paper presentation]. 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 378–384). https://doi.org/10.1109/ROMAN.2017.8172330

Rossi, A., & Rossi, S. (2021). *Engaged by a bartender robot: Recommendation and personalisation in human-robot interaction* [Paper presentation]. Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '21, New York, NY, USA (pp. 115–119). Association for Computing Machinery. https://doi.org/10.1145/3450614.3463423

Rudovic, O., Zhang, M., Schuller, B., & Picard, R. (2019). *Multi-modal active learning from human data: A deep reinforcement learning approach* [Paper presentation]. 2019 International Conference on Multimodal Interaction, ICMI '19, New York, NY, USA (pp. 6–15). Association for Computing Machinery. https://doi.org/10.1145/3340555.3353742

Salam, H., Celiktutan, O., Hupont, I., Gunes, H., & Chetouani, M. (2017). Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, *5*, 705–721. https://doi.org/10.1109/ACCESS.2016.2614525

Saleh, K., Yu, K., & Chen, F. (2021). *Improving users engagement detection using end-to-end spatio-temporal convolutional neural networks* [Paper presentation]. Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, HRI '21 Companion New York, NY, USA (pp. 190–194). Association for Computing Machinery. https://doi.org/10.1145/3434074.3447157

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). *Automatic analysis of affective postures and body motion to detect engagement with a game companion* [Paper presentation]. Proceedings of the 6th International Conference on Human-Robot Interaction, HRI '11, New York, NY, USA (pp. 305–312). Association for Computing Machinery. https://doi.org/10.1145/1957656.1957781

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Shenoy, S., Jiang, Y., Lynch, T., Manuel, L. I., & Doryab, A. (2022). *A self learning system for emotion awareness and adaptation in humanoid robots* [Paper presentation]. 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 912–919). https://doi.org/10.1109/RO-MAN53752.2022.9900581

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). *Where to look: A study of human-robot engagement* [Paper presentation]. Proceedings of the 9th International Conference on Intelligent User Interfaces, IUI '04, New York, NY, USA (pp. 78–84). Association for Computing Machinery. https://doi.org/10.1145/964456.964458

Silvera-Tawil, D., Bruck, S., Xiao, Y., & Bradford, D. (2022). Socially-assistive robots to support learning in students on the autism spectrum: Investigating educator perspectives and a pilot trial of a mobile platform to remove barriers to implementation. *Sensors*, *22*(16), 6125. https://doi.org/10.3390/s22166125

Silvia Ivani, A., Giubergia, A., Santos, L., Geminiani, A., Annunziata, S., Caglio, A., Olivieri, I., & Pedrocchi, A. (2022). A gesture recognition algorithm in a robot therapy for asd children. *Biomedical Signal Processing and Control*, *74*, 103512. https://doi.org/10.1016/j.bspc.2022.103512

Simões, G., Lopes, A., Carona, C., Pereira, R., & Nunes, U. J. (2023). *Deep-learning based classification of engagement for child-robot interaction* [Paper presentation]. 2023 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) (pp. 112–117). https://doi.org/10.1109/ICARSC58346.2023.10129551

Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2023). Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, *14*(2), 1012–1027. https://doi.org/10.1109/TAFFC.2021.3127692

Świetlicka, A., Haczyk, D., & Haczyk, M. (2023). *Graph neural networks for natural language processing in human-robot interaction* [Paper presentation]. 2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) (pp. 89–94). https://doi.org/10.23919/SPA59660.2023.10274451

Thomas, J., & Graziosi, J. B., S. (2010). *EPPI-Reviewer 4.0: Software for research synthesis. EPPI centre software. EPPI Centre Software*. Social Science Research Unit.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). Prisma extension for scoping reviews (prisma-scr): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Trinh, H., Shamekhi, A., Kimani, E., & Bickmore, T. W. (2018). *Predicting user engagement in longitudinal interventions with virtual agents* [Paper presentation]. Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA '18, New York, NY, USA (pp. 9–16). Association for Computing Machinery. https://doi.org/10.1145/3267851.3267909

Tuyen, N. T. V., Jeong, S., & Chong, N. Y. (2018). *Incremental learning of human emotional behavior for social robot emotional body expression* [Paper presentation]. 2018 15th International Conference on Ubiquitous Robots (UR) (pp. 377–382). https://doi.org/10.1109/URAI.2018.8441767

Vaufreydaz, D., Johal, W., & Combe, C. (2016). Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems*, *75*, 4–16. https://doi.org/10.1016/j.robot.2015.01.004

Weng, W.-T., Huang, H.-P., Zhao, Y.-L., & Lin, C.-Y. (2022). Development of a visual perception system on a dual-arm mobile robot for human-robot interaction. *Sensors*, *22*(23), 9545. https://doi.org/10.3390/s22239545

Weng, Y.-H., & Hirata, Y. (2022). Design-Centered HRI governance for healthcare robots. *Journal of Healthcare Engineering*, *2022*(1), 3935316. https://doi.org/10.1155/2022/3935316

Zhang, Z., Zheng, J., & Magnenat Thalmann, N. (2022). Engagement estimation of the elderly from wild multiparty human–robot interaction. *Computer Animation and Virtual Worlds*, *33*(6), e2120. https://doi.org/10.1002/cav.2120

Zhang, Z., Zheng, J., & Thalmann, N. M. (2021). *Engagement intention estimation in multiparty human-robot interaction* [Paper presentation]. 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) (pp. 117–122). https://doi.org/10.1109/RO-MAN50785.2021.9515373

About the authors

Bahram Salamat Ravandi is a PhD student in Cognitive Science. His research focuses on Human-Robot Interaction, Affective Computing, and studying engagement and human behavior analysis. He develops experimental setups with social robots for emotionally intelligent interactions and is also interested in human-autonomous vehicle interaction and explainable AI.

Imran Khan is a Marie-Sködowska Curie Actions EUTOPIA-SIF Postdoctoral Fellow at the DICE Lab (University of Gothenburg). His research focuses on areas of artificial intelligence and artificial life, affective computing, and human-computer interaction with a focus on socially assistive robots. http://www.imytk.co.uk.

Pierre Gander is an Associate Professor, senior lecturer, and head of the bachelor's program in cognitive science at the University of Gothenburg. He also leads the UNREAL international network for research on cognition and fictionality. He conducts research on cognition, imagination, episodic memory, and the reality-fiction distinction in memory.

Robert Lowe is an Associate Professor at the University of Gothenburg and Senior Researcher at RISE AB. He conducts research in Human-Technology Interaction with a focus on digitized cognitive training, Human-Robot Interaction, and Driver support systems in relation to the safe use of AI and monitoring cognitive-affective states. http://dice-r-lab.com/robert.