Quantifying User Engagement in a Triadic Human-Robot Interaction Setup: Incorporating Gaze, Head Pose, and Affective Cues

Bahram Salamat Ravandi[1,*], John Currie[1], Pierre Gander[1] and Robert Lowe[1]

[1]Forskningsgången 6 - 417 56, Gothenburg, Sweden, Department of Applied IT, University of Gothenburg

Abstract

Recent research in Human-Robot Interaction (HRI) has increasingly focused on understanding user engagement to enhance the overall user experience. This paper aims to develop a predictive model of user engagement within a triadic interaction loop involving three key entities: a human, a robot, and a task. To achieve this, we created a new dataset incorporating multimodal features, including facial landmarks, facial action units, head posture, and gaze. Engagement annotations were performed by two human annotators using a structured approach to ensure high-quality labeling. Building upon this dataset, we developed a deep learning-based predictive model of user engagement. The results demonstrate that the model effectively captures user engagement in the task-oriented HRI scenario, achieving a Mean Squared Error (MSE) of 0.0111 and an R^2 score of 0.8195, highlighting its accuracy and robustness. Additionally, a permutation feature importance analysis revealed that gaze, head pose, and facial expressions significantly contributed to the model's predictions across various levels of user engagement.

Keywords

Engagement, Socially Assistive Robots, Affective Engagement, Social Engagement

1. Introduction

The study of engagement has emerged as a response to the desire to create services, products, and content that are tailored to user experience in order to engage users [1]. In the field of Human-Robot Interaction (HRI), engagement is a multifaceted concept with diverse definitions in the literature [1]. Due to the diverse ways engagement is understood in the HRI field, researchers have employed a wide range of metrics and features to measure it [1, 43]. In this paper, we define 'engagement' as:

> a quality of user experiences with technology that is characterized by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control

SAIS2025: Swedish AI Society Workshop 2025, 16-17 June 2025, Halmstad, Sweden.

*Corresponding author.

bahramsalamat@ait.gu.se (B. S. Ravandi); john.currie@ait.gu.se (J. Currie); pierre.gander@ait.gu.se (P. Gander); robertlowe@ait.gu.se (R. Lowe)

https://www.linkedin.com/in/bahramsalamat/ (B. S. Ravandi); https://www.linkedin.com/in/pierre-gander-b76353292 (P. Gander); https://www.linkedin.com/in/robert-lowe-02836b4 (R. Lowe)

0009-0008-2712-8684 (B. S. Ravandi); 0009-0004-9920-3680 (J. Currie); 0000-0002-0214-7511 (P. Gander); 0000-0002-0307-3171 (R. Lowe)

and time, awareness, motivation, interest, and affect ([11], p. 949).

One prominent area where engagement plays a critical role is Socially Assistive Robots (SARs), which have been increasingly deployed in fields such as education and healthcare. For instance, [2] developed two Machine Learning (ML) models to monitor long-term engagement with SARs, particularly for children with autism spectrum disorder. By utilizing audio-visual and performance data, they trained several ML algorithms and implemented re-engagement strategies when engagement levels dropped below a certain threshold. Similarly, [3] proposed an assistive robot designed to support Alzheimer's patients through memory training exercises, leveraging verbal and nonverbal communication to sustain user engagement. They identified four distinct levels of engagement that robots must adapt to based on user performance, ensuring a dynamic and responsive interaction.

Various Machine Learning and Deep Learning models have been applied to different datasets to improve accuracy and adaptability in engagement detection [39]. For instance, [17] used the UE-HRI dataset to develop a 3D Convolutional Neural Network (CNN) model that detects engagement based on video frame sequences. Similarly, [18] built a CNN model to identify passive subjects in a four-way HRI interaction using facial and speech data. Other studies have explored alternative approaches. For example, [19] utilized a Long Short-Term Memory (LSTM) model on visual data, body pose, and facial features to detect disengagement in children with learning difficulties, while [20] applied a Recurrent Neural Network (RNN) to model engagement using behavioral and speech data from the UE-HRI dataset. Expanding on these methods, [2] compared different ML algorithms for engagement detection based on facial, audio, and game performance features. [2] evaluated several conventional model types, including Naïve Bayes, K-nearest neighbors, support vector machines, neural networks, logistic regression, random decision tree forests, and gradient-boosted decision trees. Among these, gradient-boosted decision trees emerged as the most successful, achieving the highest Area Under the Receiver Operating Characteristic (AUROC) values. Further, [15] implemented a multimodal active learning approach with Reinforcement Learning (RL) and LSTMs to detect child-robot engagement, while [21] trained CNN and LSTM models to classify different engagement levels in interactions with the TEGA robot. Other works focused on personalization and multimodal engagement detection. [22] introduced CultureNet, a CNN-based model for personalized engagement detection, while [24] combined CNN models for facial expression and body posture analysis to classify engagement into positive, negative, and neutral categories. Additionally, [9] trained an RNN model on the EASE dataset, incorporating videos, audio, and physiological signals, using facial action units as input features. Finally, [25] developed a Multi-Task Cascaded Convolutional Neural Networks (MTCNN) model using facial landmarks and Histogram of Oriented Gradients (HOG) features to detect engagement states in children.

However, these models necessitate substantial amounts of training data, and there are only a limited number of existing engagement datasets. Several datasets have been used in HRI for engagement detection [39]. One well-known dataset is the UE-HRI dataset [4, 5], which was collected from human interactions with the Pepper robot in a public space. This dataset includes video, voice, sonar, and laser data, with engagement annotations primarily focused on cues related to disengagement. A notable limitation of the UE-HRI dataset is that it focuses solely on the binary presence or absence of disengagement, without capturing the full spectrum

or intensity of user engagement and emotional states. This limitation could hinder the model's ability to perceive more nuanced engagement levels during interactions.

Another significant dataset is the TOGURO dataset, gathered from human interactions with the NAO robot in public settings [6, 7]. It contains video streams, as well as verbal and non-verbal user behaviors, along with user position data. In addition to these engagement-specific datasets, several emotion-based datasets are commonly used in research, including the Static Facial Expressions in the Wild (SFEW), Facial Expression Recognition (FER2013), and AffectNet [39]. However, these datasets rely heavily on facial expressions to detect affective engagement and overlook other significant indicators of user engagement, such as pose and gaze.

Due to the subjective and context-dependent nature of engagement, annotating engagement is both time-consuming and challenging. While engagement annotation is typically performed manually, alternative approaches have been explored. For instance, [9] combined self-reports with expert annotations to establish ground truth, whereas [10] employed unsupervised methods to categorize engagement into four patterns: approaching, interacting, leaving, and uninterested. Following the establishment of this structured engagement dataset, a deep neural network model was trained to detect user engagement.

In this paper, we introduce an engagement dataset and engagement predictive model within a triadic human-robot-task interaction. We constructed a dataset and implemented a rigorous engagement annotation methodology to guarantee high-quality data labeling. The annotation was informed by insights gained from data collection in [53], where observing patterns of user behavior in interaction videos illuminated the correlations between various indicators (such as facial expressions and gaze behavior) and engagement levels. The developed dataset incorporates a variety of multimodal features, including facial landmarks, head pose, and gaze direction.

The following sections will detail the proposed general HRI setup, engagement annotation, and modeling methodology.

2. HRI Framework

This study aims to assess human engagement in a triadic HRI setup by introducing an engagement annotation framework and developing an engagement predictive model. This model can potentially be used to enhance user experience by re-engaging users or increasing user engagement via social and instructional feedback from the robot or by dynamically adjusting task difficulty. Various gamification elements can be integrated into the setup, allowing users to engage with the task through rewards and audiovisual feedback from both the robot and the task itself [16]. The feasibility of implementing such a function depends on real-time engagement assessment.

The interaction loop, as presented in [16], consists of six components:

1. **Challenge Modulation:** Adjusting the task's difficulty based on the user's engagement state. If disengagement arises due to excessive difficulty, reducing the challenge can help re-engage the user. Conversely, if the task is too easy, increasing its difficulty may enhance engagement by offering a more stimulating experience. According to Flow theory, an

individual can experience a state of deep satisfaction and immersion when there is an optimal balance between the task's challenge and their skill level [32].

2. **Task State:** This component involves providing information on human performance considering the current state of the task. It functions as a gamification element, tracking and displaying user progress and achievements, which can motivate sustained engagement and improvement [30, 31, 16].
3. **Action Selection:** This component involves the use of touchscreen-based, verbal, or mouse inputs for selecting actions. Providing users with hints, tips, or instructions as a gamification element can result in higher engagement and improve users' performance [16, 41, 40].
4. **Reward Feedback:** The task provides direct feedback on the outcome of a specific action taken by the user. This module is considered as a within-task gamification element in an HRI setup [39].
5. **Social Feedback:** This component includes the robot's verbal/nonverbal feedback to encourage desired behaviors and acknowledge accomplishments [33, 34, 35, 36]. An engagement assessment model can assist in determining appropriate robot responses. Performance-based feedback may be particularly beneficial for tasks focused on achievement, such as cognitive training or educational activities. Conversely, in scenarios where fostering social connections is essential, such as companionship or social skills development, emphasizing affective-based feedback can enhance user engagement with the robot.
6. **Engagement State:** This component refers to possible inputs that can help determine the user's emotional state or level of engagement. These features may include facial expressions, physiological signals (e.g., EEG, GSR, ECG), eye-tracking data, and body movements captured by Kinect sensors.

To effectively engage users, feedback and task difficulty adjustments must be adapted to the user's engagement state, estimated using an engagement estimation model. The literature presents various adaptive strategies based on engagement, such as rule-based systems that adjust according to user engagement levels [38, 23, 37], and Reinforcement Learning (RL)-based policy learners, enabling more tailored adaptations for user engagement [6, 2, 39].

Research Questions

1. How can we effectively quantify users' task engagement in a triadic HRI scenario involving task-oriented interactions?
2. What multimodal features (e.g., Facial Action Units (FAUs), head postures, and gaze directions) contribute to the assessment of user engagement?

3. Methodology

Fig. 1 presents the HRI setup designed for data collection, as originally introduced in [53]. The dataset consists of video recordings capturing user interactions with a social robot, Furhat, during

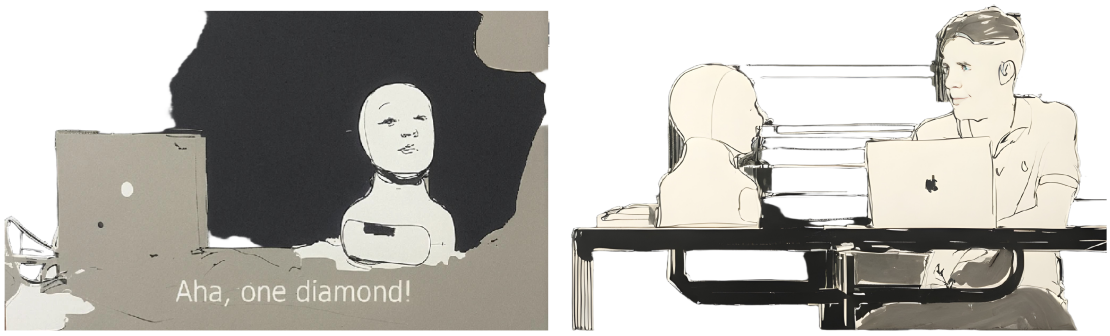

Figure 1: Interaction Loop: The user, robot (Furhat), and the gamified task are three key entities of the proposed setup. The user can select an action and receive audio-visual feedback from both the task and the robot [53].

a memory training task. The robot is positioned at an angled distance from both the user and the screen. Each interaction lasts approximately 10–15 minutes, during which the robot provides feedback aligned with the task's outcome. Additionally, the robot establishes eye contact with the user after delivering feedback in response to the user's actions. Fifty-eight engineering students (35 males, 23 females) from Koç University's Electrical and Electronics Engineering and Computer Engineering departments, aged 18 to 24 (M = 20, SD = 1.87), participated in the data collection.

3.1. Engagement Annotation

In this setup, user task engagement can be defined through indicators, such as head and gaze orientations, as well as facial expressions. A user's **head or gaze orientation** serves as an indicator of attention and engagement. When a user shifts their head or gaze away from the screen, it suggests a decline in engagement. However, even when users maintain their gaze on the screen, their level of engagement may vary between **positive, negative, or neutral states**. Through qualitative analysis of the video data, seven distinct levels of user engagement were established, which are categorized as follows:

- **Level 1**: Completely disengaged (looking away from the screen).
- **Level 2**: Occasional glances at the screen, lacking sustained focus.
- **Level 3**: User maintains attention on the screen but exhibits signs of distraction.
- **Level 4**: User maintains a steady focus on the screen while showing negative expressions (e.g., frustration or disinterest).
- **Level 5**: User maintains a steady focus on the screen and shows neutral expressions (neither positive nor negative).
- **Level 6**: User maintains a steady focus on the screen and shows positive expressions (e.g., happiness or interest).
- **Level 7**: User maintains a steady focus on the screen and is highly engaged (displaying strong positive emotions).

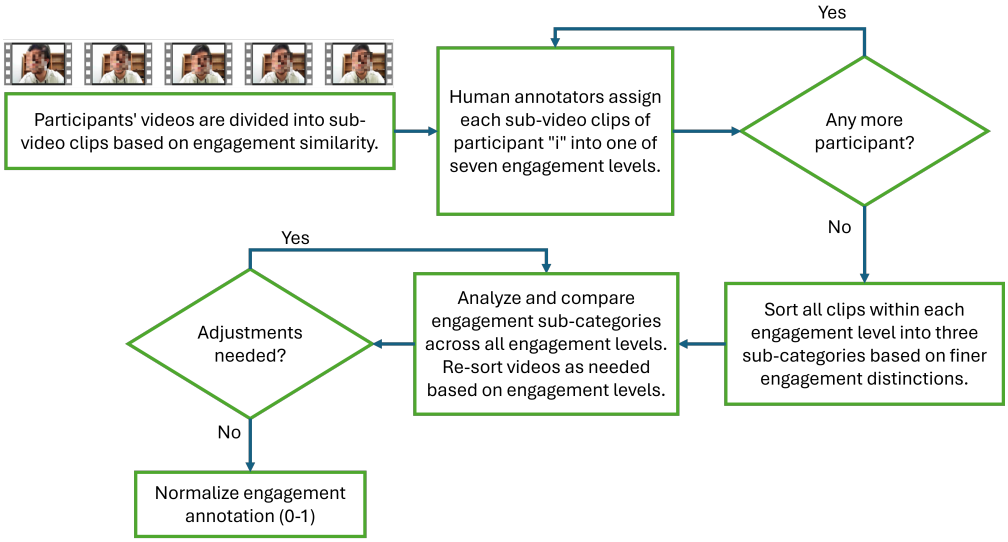

Figure 2: Annotation process: Videos are split into sub-clips, which are then categorized by engagement levels. Finally, engagement levels are refined and normalized to a 0-1 scale.

Fig. 2 illustrates a flowchart of the annotation process. The participants' videos were segmented into sub-video clips based on similar engagement patterns, which were assessed manually by a researcher. The researcher evaluated similarity by observing behavioral cues such as facial expressions, body language, gaze direction, and vocal tone, identifying segments that reflected consistent levels or patterns of engagement. This segmentation facilitates a detailed temporal analysis of user engagement by capturing variations in attention and affective states throughout the interaction. While this method may introduce a certain degree of noise or inaccuracies, given the impracticality of labeling every video frame, it enables the extraction of meaningful insights from the temporal patterns.

Two researchers conducted the labeling independently. Firstly, they conducted an initial round of annotation in which they sorted out the segmented video clips of each participant into one of the seven predefined engagement levels. This initial labeling process is essential for establishing reliable baseline engagement data. After initial labeling, clips within each engagement level were further sorted out into three subcategories to increase annotation granularity. The subcategories established within each engagement level were designed to represent different degrees of engagement, essentially dividing the levels into three further gradations, ranging from low to high within that engagement category. This approach allows for a more nuanced understanding of user engagement by capturing subtle variations in user behavior and emotional responses. To ensure robustness, annotators reviewed and compared engagement sub-categories across all levels and resorted clips if necessary.

Each annotator labeled 233 video clips, featuring interactions from 58 different participants. To assess inter-rater reliability, we calculated a weighted Cohen's Kappa coefficient [42, 44], a statistical measure that accounts for both agreement and the likelihood of chance agreement. Given that the annotation labels are ordinal, implying a meaningful order among categories, the standard Cohen's Kappa is not ideal, as it treats all disagreements equally. Instead, we apply

quadratic weighting, which penalizes larger disagreements more heavily than minor ones. This ensures that a disagreement between adjacent categories (e.g., 3 vs. 4) is considered less severe than one between distant categories (e.g., 1 vs. 5). By using weighted Kappa, we obtain a more accurate measure of inter-rater agreement that properly reflects the structure of our data. The final labels were determined by averaging the annotators' ratings.

3.2. Engagement Modeling

Engagement labels for model training were generated by averaging the scores provided by the two annotators. These averaged values were then normalized to a continuous scale ranging from 0 (indicating complete disengagement) to 1 (indicating high engagement). To extract relevant behavioral features, we used the OpenFace toolkit [12], which provides a comprehensive set of facial and gaze-related data. Specifically, features extracted include facial landmarks (2D and 3D), head pose, eye gaze, and Facial Action Units (FAUs).

The resulting dataset comprised 704 features categorized into several groups: gaze data (8), eye landmarks (168), 3D head pose (6), 2D facial landmarks (136), 3D facial landmarks (204), head pose model parameters (6), Point Distribution Model (PDM) parameters (34), FAUs (35), 3D landmark Z-coordinates (Z_0 to Z_{67}), and 39 AU-related parameters. The complete dataset is publicly available at https://osf.io/4nfwh. All feature values were standardized prior to training to ensure uniformity across scales.

The predictive model was developed using a deep learning architecture implemented in `TensorFlow` and `Keras`. It consists of fully connected layers utilizing ReLU activation functions and dropout regularization to mitigate overfitting. The model architecture is summarized in Fig. 3. Since the target variable represents a probability of user engagement, a sigmoid activation function is used in the output layer to ensure predictions remain within the $[0, 1]$ range.

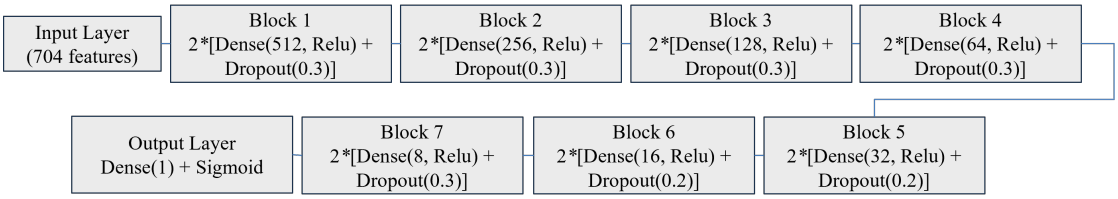

Figure 3: Model architecture summary

To evaluate model performance, the dataset was partitioned into training (80%) and testing (20%) subsets. To maintain a balanced representation of engagement levels and participant data, training samples were randomly selected across participants and engagement levels. This strategy helps to prevent bias due to over-representation of specific cases and promotes generalizability. To further enhance model robustness and increase data variability, we employed data augmentation by generating vertically mirrored versions of the video clips. These augmented clips were assigned the same engagement labels as their original counterparts. Model optimization was performed using the Adam optimizer with Mean Squared Error (MSE) as the loss function.

4. Results

The inter-rater agreement analysis yielded a weighted Cohen's Kappa of 0.91, indicating a high level of agreement between the annotators. The predictive model's performance was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Table 1 summarizes the final evaluation outcomes for the model. To better understand performance variations across engagement levels, we evaluated the model separately on test subsets corresponding to each engagement level. This allowed us to assess how well the model generalizes across different engagement levels, despite being trained holistically. Fig. 4 displays the loss curves corresponding to four batch sizes (30, 40, 50, and 60), illustrating the model's fine-tuning process. The batch size of around 40 yields the most stable performance, suggesting it is well-suited to the dataset's characteristics and the labeling approach used.

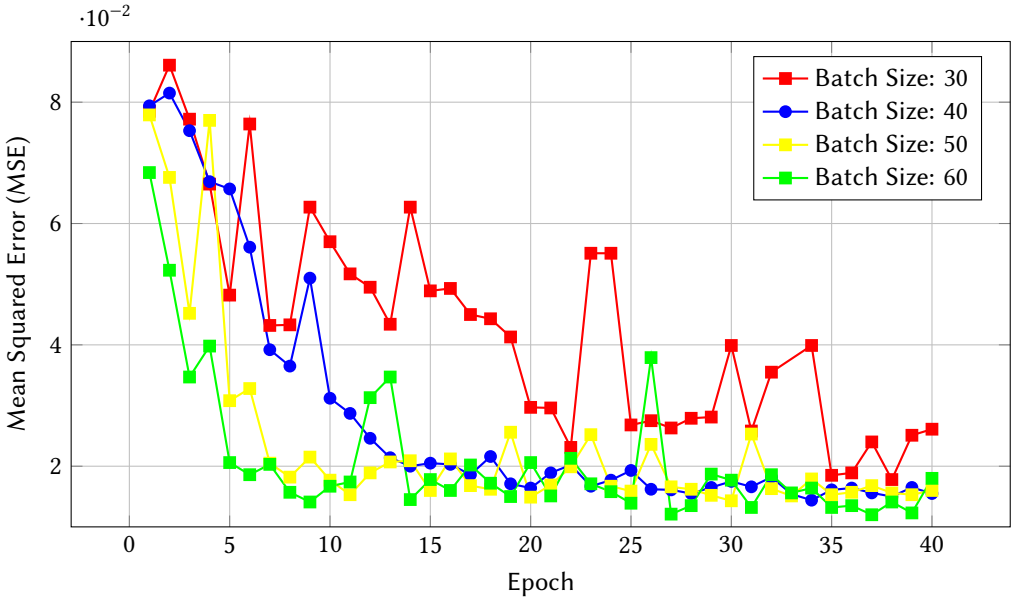

Figure 4: Comparison of MSE per epoch for batch sizes: 30, 40, 50 and 60

Engagement Level	MSE	MAE	RMSE	R^2
1	0.0039	0.0361	0.0627	-
2	0.0781	0.2410	0.2794	-
3	0.1211	0.3335	0.3480	-
4	0.0222	0.1220	0.1489	-
5	0.0036	0.0463	0.0601	-
6	0.0042	0.0409	0.0648	-
7	0.0458	0.2111	0.2139	-
Total	0.0111	0.0696	0.1054	0.8195

Table 1
Evaluation metrics by engagement levels

The model achieved an MSE of 0.0111, indicating low squared error on average across predictions. The MAE was 0.0696, suggesting an average absolute deviation of approximately 6.96%. Notably, the R^2 score of 0.8195 indicates that the model accounts for 81.95% of the variance in engagement values, demonstrating strong predictive performance. These results suggest that the model effectively captures the underlying engagement patterns. However, performance was comparatively lower for engagement levels 2, 3, and 7, likely due to the limited amount of training data available for these categories.

To better understand the relationship between the specific features of engagement and the model prediction of engagement level, a feature importance analysis was conducted. Table 2 presents the most influential features ranked by their **permutation importance** values, derived from the trained model based on MSE as a performance metric. Permutation importance measures the decrease in a model's performance when the values of a single feature are randomly shuffled. A higher value indicates a greater contribution of the feature to the model's predictions [47]. These values were calculated using the trained model, with MSE as the evaluation metric. Notably, a negative permutation importance suggests that scrambling the feature improves model performance, potentially indicating overfitting, where the model relies on misleading patterns not generalizable to new data. For each engagement level, the six most important features are listed. This table provides a more detailed analysis of how the trained model differentiates between engagement levels. Fig. 5 presents selected frames from a video in the dataset, showing a participant interacting with both the task and the robot. Each frame includes the predicted engagement value. Lower predicted values correspond to lower engagement levels, as detailed in Table 2. For instance, frame 3 has a predicted engagement score of 0.61 and is associated with facial expressions such as smiling and raised cheeks.

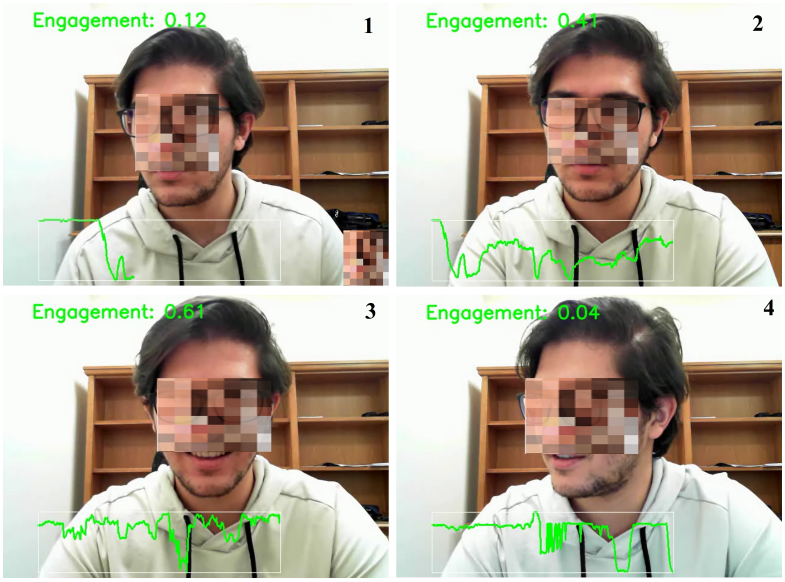

Figure 5: Several frames from a video in the dataset recorded from the interaction of a participant with the task and the robot. Facial regions have been blurred for anonymization. The green plot illustrates engagement value changes over the sequence frames.

Engagement Level	Top Features (Feature = Importance)
Level 7	p_23 = 0.000486, AU23_r = 0.000286, p_29= 0.000157, AU15_c = 0.000137, AU20_c = 0.000130, AU23_c = 0.000111
Level 6	AU10_c = 0.000387, AU12_c = 0.000239, AU12_r = 0.000201, p_0 = 0.000103, AU04_r = 0.000089, p_30 = 0.000085
Level 5	p_18 = 0.000113, AU07_c = 0.000091, p_16 = 0.000085, AU09_c = 0.000084, X_3 = 0.000078, X_0 = 0.000076
Level 4	p_19 = 0.000698, p_17 = 0.000637, p_4 = 0.000478, p_18 = 0.000379, pose_Rz = 0.000354, p_rz = 0.000344
Level 3	gaze_0_x = 0.003535, gaze_angle_x = 0.003366, gaze_1_x = 0.003304, pose_Ry = 0.003282, p_ry = 0.002919, p_1 = 0.002722
Level 2	gaze_angle_x = 0.032542, gaze_1_x = 0.032406, gaze_0_x = 0.029830, pose_Ry = 0.019259, p_ry = 0.016968, p_1 = 0.010558
Level 1	gaze_angle_x = 0.007903, gaze_0_x = 0.007356, gaze_1_x = 0.006444, pose_Ry = 0.003462, p_1 = 0.002450, p_ry = 0.002113
Total	gaze_angle_x = 0.001091, gaze_0_x = 0.001091, gaze_1_x = 0.000949, pose_Ry = 0.000560, p_ry = 0.000381, p_1 = 0.000340

Table 2
Top features for each engagement level

- Gaze and pose features ranked highest in importance in engagement levels 1, 2, and 3, indicating that gaze direction, orientation, and pose are highly informative for the prediction task in these levels. Gaze and body orientation are often key signals in assessing engagement and attention [50, 51, 5].
- Facial Action Units show higher importance in levels 5, 6, and 7, suggesting that expressions play a significant role in the model's decision-making process in these levels. In the literature, facial expressions like smiles and raised cheeks are associated with positive emotions [23, 24].
- Facial landmark features, while less dominant individually, still contributed meaningfully. While facial landmarks are not the most influential on their own, they could still play a key role when considered in conjunction with other features, as discussed in studies on the integration of multiple engagement cues [39, 52].

5. Conclusion

This study presents a predictive model for assessing user engagement in a triadic human-robot interaction setup, consisting of a human, a robot, and a task. The model is built using a novel

dataset that we developed, incorporating multimodal features such as facial landmarks, facial action units, head pose, and gaze direction. Additionally, we introduce a structured framework for annotating engagement, addressing a significant gap in existing research on systematic engagement annotation in the HRI scenario.

Engagement annotations were carried out using a structured approach, resulting in a weighted Cohen's Kappa score of 0.91, reflecting a high level of agreement among the annotators. The predictive model showed excellent performance, with a Mean Squared Error (MSE) of 0.0111 and an R^2 score of 0.8195. These results demonstrate the model's ability to accurately capture user engagement patterns, suggesting its potential for adapting real-time interactions based on engagement states.

Engagement is a complex phenomenon that cannot be fully understood through a single modality. The model's differential weighting of various features supports the idea that engagement detection benefits from a multimodal approach. For instance, gaze orientation, a major feature of attention, is closely tied to social presence in robotic companions. In this study, this correlates with the lower engagement level where the users spend time looking away from the screen or making eye contact with the robot. Head pose, another key indicator, reflects body language and attentiveness. This suggests that in adaptive HRI, robots capable of interpreting users' head pose could tailor their responses or adjust task difficulty in real-time to re-engage users, highlighting the potential for robotic systems to leverage machine learning models to assess and respond to user attentiveness, beyond just task performance. Furthermore, the model's reliance on facial action units to detect higher levels of engagement points to a crucial intersection between facial action units and engagement. For example, robots that adjust their behavior based on positive emotional cues, like smiles, could enhance user satisfaction and prolong engagement.

Although task parameters were not explicitly included as variables in the analysis, all behavioral data — including facial expressions, gaze, and head pose — were collected during a structured visuospatial memory task. As such, these cues are inherently tied to participants' engagement with the task. Given the cognitive demands of the activity, the model is likely to generalize well to other high-tempo cognitive scenarios, such as video games, cognitive training programs, or driving simulations.

Despite these contributions, the study acknowledges the importance of context in evaluating engagement, noting that engagement is context-sensitive and can vary across tasks. For example, placing the robot beyond the screen, could affect engagement outcomes [49], resulting in less attention directed at the robot if it is placed in the peripheral vision. In different tasks, whether non-social or other types of social interactions, users may express their engagement in distinct ways [48]. This distinction underscores the task-dependent nature of affective states and engagement. We acknowledge that the use of facial expression, head pose, and gaze orientation captures some aspects of user engagement — primarily affect, attention, and interest — but misses many of the cognitive, behavioral, and experiential dimensions outlined in the broader definition. Furthermore, there is a need for further research to broaden the generalizability of these findings by incorporating diverse user populations. Future studies should aim to incorporate user experiences and physiological data to gain deeper insights into affective states and enhance the model's reliability.

Moreover, the permutation feature importance analysis is sensitive to feature collinearity

and dependent on a single trained model. While more advanced methods account for feature interactions and offer uncertainty bounds on feature importance, they were beyond the specific objectives of this study, which aimed to provide an initial understanding of the relative importance of engagement features. For example, Fisher et al. [46] introduced a framework that evaluates feature importance across the entire class of well-performing models, known as Model Class Reliance. This provides bounds on a feature's importance and accounts for feature interactions and redundancy. Similarly, SHAP (SHapley Additive exPlanations) offers an explanation method grounded in cooperative game theory, attributing contributions to individual features while accounting for interactions [45].

In conclusion, this research establishes a foundation framework for understanding and quantifying user engagement in HRI, presenting significant advancements and practical implications while also identifying critical areas for further investigation.

Declaration on Generative AI

During the preparation of this manuscript, the author(s) utilized ChatGPT-4 solely for the purpose of grammar and spelling verification. No text was generated by the generative AI, and the author(s) take(s) full responsibility for the content of this publication. The author(s) used (https://photo-to-sketch.ai) for converting Figure 1 into sketch-style illustrations, and no further alterations were made.

References

[1] K. Doherty and G. Doherty, "Engagement in HCI: Conception, Theory and Measurement," ACM Computing Surveys, vol. 51, no. 5, pp. 1–39, 2019.

[2] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders," Science Robotics, vol. 5, no. 39, 2020.

[3] A. Andriella, C. Torras, and G. Alenyà, "Cognitive System Framework for brain-training exercise based on human-robot interaction," Cognitive Computation, vol. 12, no. 4, pp. 793–810, 2020.

[4] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions," Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 464-472, January 2017.

[5] A. Ben-Youssef, G. Varni, S. Essid, and C. Clavel, 'On-the-Fly Detection of User Engagement Decrease in Spontaneous Human–Robot Interaction Using Recurrent and Deep Neural Networks', International Journal of Social Robotics, vol. 11, no. 5, pp. 815–828, Dec. 2019.

[6] F. del Duchetto, P. Baxter, and M. Hanheide, "Automatic assessment and learning of robot social abilities," ACM/IEEE International Conference on Human-Robot Interaction, pp. 561–563, 2020.

[7] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous en-

gagement assessment from a robot's point of view," Frontiers in Robotics and AI, vol. 7, 2020.

[8] N. Poltorak and A. Drimus, "Human-robot interaction assessment using dynamic engagement profiles," IEEE-RAS International Conference on Humanoid Robots, pp. 649–654, 2017.

[9] S. Dhamija and T. E. Boult, "Automated Action Units Vs. Expert Raters: Face off," Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, vol. 2018-Janua, pp. 259–268, 2018.

[10] S. P. Pattar, E. Coronado, L. R. Ardila, and G. Venture, "Intention and engagement recognition for personalized human-robot interaction, an integrated and deep learning approach," 2019 4th IEEE International Conference on Advanced Robotics and Mechatronics, ICARM 2019, pp. 93–98, 2019.

[11] O'Brien, H. & Toms, E. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal Of The American Society For Information Science And Technology*. **59**, 938-955 (2008).

[12] Baltrušaitis, T., Robinson, P. & Morency, L. OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference On Applications Of Computer Vision (WACV)*. pp. 1-10 (2016)

[13] I. Goodfellow, D. Erhan, PL. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, DH. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio "Challenges in Representation Learning: A report on three machine learning contests," International Conference on Machine Learning (ICML) 2013.

[14] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," Sci Robot, vol. 3, no. 19, 2018.

[15] O. Rudovic, M. Zhang, B. Schuller, and R. W. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction, pp. 6–15, 2019.

[16] Ravandi, B. Gamification for Personalized Human-Robot Interaction in Companion Social Robots. *2024 12th International Conference On Affective Computing And Intelligent Interaction Workshops And Demos (ACIIW)*. pp. 106-110 (2024).

[17] K. Saleh, K. Yu, and F. Chen, "Improving users engagement detection using end-to-end spatio-temporal convolutional neural networks," Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, 2021.

[18] D. Ayllon, T.-S. Chou, A. King, and Y. Shen, "Identification and engagement of passive subjects in multiparty conversations by a humanoid robot," Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, 2021.

[19] G. K. Sidiropoulos, G. A. Papakostas, C. Lytridis, C. Bazinas, V. G. Kaburlasos, E. Kourampa, and E. Karageorgiou, "Measuring engagement level in child-robot interaction using machine learning based data analysis," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020.

[20] A. Atamna and C. Clavel, "HRI-RNN: A user-robot dynamics-oriented RNN for engagement decrease detection," Interspeech 2020, 2020.

[21] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[22] O. Rudovic, Y. Utsumi, J. Lee, J. Hernandez, E. C. Ferrer, B. Schuller, and R. W. Picard, "Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018.

[23] A. Mollahosseini, H. Abdollahi, and M. H. Mahoor, "Studying Effects of Incorporating Automated Affect Perception with Spoken Dialog in Social Robots," RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication, pp. 783–789, 2018.

[24] A. Rajavenkatanarayanan, A. R. Babu, K. Tsiakas, and F. Makedon, "Monitoring task engagement using facial expressions and body postures," Proceedings of the 3rd International Workshop on Interactive and Spatial Computing - IWISC '18, 2018.

[25] A. Di Nuovo, D. Conti, G. Trubia, S. Buono, and S. Di Nuovo, "Deep Learning Systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability," Robotics, vol. 7, no. 2, p. 25, 2018.

[26] D. Anagnostopoulou, N. Efthymiou, C. Papailiou, and P. Maragos, "Engagement Estimation During Child Robot Interaction Using Deep Convolutional Networks Focusing on ASD Children," no. June, pp. 3641–3647, 2021.

[27] R. Garris, R. Ahlers, and J. E. Driskell, "Games, motivation, and learning: A research and Practice Model," Simulation & Gaming, vol. 33, no. 4, pp. 441–467, 2002.

[28] T. Alves, S. Gama, and F. S. Melo, "Flow adaptation in serious games for health," 2018 IEEE 6th International Conference on Serious Games and Applications for Health, SeGAH 2018, pp. 1–8, 2018.

[29] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 10, pp. 1943–1955, 2016.

[30] Ahmad, M., Mubin, O. & Orlando, J. Children views' on social robot's adaptations in education. *Proceedings Of The 28th Australian Conference On Computer-Human Interaction*. pp. 145-149 (2016), https://doi.org/10.1145/3010915.3010977.

[31] Liles, K. Ms. An (Meeting Students' Academic Needs): Engaging Students in Math Education. *Adaptive Instructional Systems*. pp. 645-661 (2019).

[32] Csikszentmihalyi, M. Finding flow: The psychology of engagement with everyday life.. (Basic Books,1997).

[33] Brown, L., Kerwin, R. & Howard, A. Applying Behavioral Strategies for Student Engagement Using a Robotic Educational Agent. *2013 IEEE International Conference On Systems, Man, And Cybernetics*. pp. 4360-4365 (2013).

[34] Boccanfuso, L., Wang, Q., Leite, I., Li, B., Torres, C., Chen, L., Salomons, N., Foster, C., Barney, E., Ahn, Y., Scassellati, B. & Shic, F. A thermal emotion classifier for improved human-robot interaction. *2016 25th IEEE International Symposium On Robot And Human Interactive Communication (RO-MAN)*. pp. 718-723 (2016).

[35] Javed, H. & Park, C. Interactions With an Empathetic Agent: Regulating Emotions and

Improving Engagement in Autism. *IEEE Robot Autom Mag.* **26**, 40-48 (2019,4).
[36] Amanatiadis, A., Kaburlasos, V., Dardani, C. & Chatzichristofis, S. Interactive social robots in special education. *2017 IEEE 7th International Conference On Consumer Electronics - Berlin (ICCE-Berlin).* pp. 126-129 (2017).
[37] Abdelrahman, A., Strazdas, D., Khalifa, A., Hintz, J., Hempel, T. & Al-Hamadi, A. Multimodal Engagement Prediction in Multiperson Human–Robot Interaction. *IEEE Access.* **10** pp. 61980-61991 (2022).
[38] Duque-Domingo, J., Gómez-García-Bermejo, J. & Zalama, E. Gaze control of a robotic head for realistic interaction with humans. *Frontiers In Neurorobotics.* **14** (2020).
[39] Ravandi, B., Khan, I., Gander, P. & And, Lowe, R. Deep Learning Approaches for User Engagement Detection in Human-Robot Interaction: A Scoping Review. *International Journal Of Human–Computer Interaction.* pp. 1-19 (2025), https://-doi.org/10.1080/10447318.2025.2470277.
[40] Arshad, N., Hashim, A., Mohd Ariffin, M., Mohd Aszemi, N., Low, H. & Norman, A. Robots as Assistive Technology Tools to Enhance Cognitive Abilities and Foster Valuable Learning Experiences among Young Children With Autism Spectrum Disorder. *IEEE Access.* **8** pp. 116279-116291 (2020).
[41] Manh Do, H., Sheng, W., Harrington, E. & Bishop, A. Clinical Screening Interview Using a Social Robot for Geriatric Care. *IEEE Transactions On Automation Science And Engineering.* **18**, 1229-1242 (2021).
[42] McHugh, M. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* **22**, 276-282 (2012).
[43] Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C. & Peters, C. Engagement in human-agent interaction: An overview. *Frontiers In Robotics And AI.*
[44] Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin.* **70**, 213 (1968).
[45] Lundberg, S. & Lee, S. A unified approach to interpreting model predictions. *Advances In Neural Information Processing Systems.* **30** (2017.
[46] Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal Of Machine Learning Research.* **20**, 1-81 (2019).
[47] Breiman, L. Random forests. *Machine Learning.* **45** pp. 5-32 (2001).
[48] Borges, N., Lindblom, L., Clarke, B., Gander, A. & Lowe, R. Classifying confusion: autodetection of communicative misunderstandings using facial action units. *2019 8th International Conference On Affective Computing And Intelligent Interaction Workshops And Demos (ACIIW).* pp. 401-406 (2019).
[49] Markelius, A., Sjöberg, S., Bergström, M., Ravandi, B., Vivas, A., Khan, I. & Lowe, R. Differential Outcomes Training of Visuospatial Memory: A Gamified Approach Using a Socially Assistive Robot. *International Journal Of Social Robotics.* **16**, 363-384 (2024,2), https://doi.org/10.1007/s12369-023-01083-0.
[50] Hadfield, J., Chalvatzaki, G., Koutras, P., Khamassi, M., Tzafestas, C. & Maragos, P. A Deep Learning Approach for Multi-View Engagement Estimation of Children in a Child-Robot Joint Attention Task. *2019 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS).* pp. 1251-1256 (2019).

[51] Rossi, A., Raiano, M. & Rossi, S. Affective, cognitive and behavioural engagement detection for human-robot interaction in a bartending scenario. *2021 30th IEEE International Conference On Robot & Human Interactive Communication (RO-MAN)*. pp. 208-213 (2021).

[52] Bartlett, M., Stewart, T. & Thill, S. Estimating levels of engagement for social human-robot interaction using legendre memory units. *Companion Of The 2021 ACM/IEEE International Conference On Human-Robot Interaction*. pp. 362-366 (2021).

[53] Ravandi, B. S., Khan, I., Markelius, A., Bergström, M., Gander, P., Erzin, E., & Lowe, R. Exploring Task and Social Engagement in Companion Social Robots: A Comparative Analysis of Feedback Types. *Manuscript in review*.