International Journal of Robotics and Control Systems
Vol. 5, No. 4, 2025, pp. 2265-2289
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

Integrating Multimodal Emotion Recognition with Deep Q-Learning for Adaptive Social Robot Interaction

Nada Khalil Al-Okbi [a,1], Saleh Ali Alomari [b,2], Raed Abu Zitar [c,3], Aseel Smerat [d,e,4], Muhannad Akram Nazzal [f,5,*], Laith Abualigah [g,6,*]

[a] Department of Computer Science, College of Science for Women, Baghdad University, Bagdad, 10070, Iraq
[b] Computer Science Department, Faculty of Information Technology, Jadara University, 21110 Jordan
[c] College of Engineering and Computing, Liwa University, Abu Dhabi, United Arab Emirates
[d] Faculty of Educational Sciences, Al-Ahliyya Amman University, Amman 19328, Jordan
[e] Centre for Research Impact and Outcome, Chitkara University, Punjab, India
[f] Faculty of Economics and Administrative Sciences, Al Albayt University, Mafraq, Jordan
[g] Department of Computer Science, Al al-Bayt University, Mafraq 25113, Jordan
[1] nada.k@csw.uobaghdad.edu.iq; [2] omari08@jadara.edu.jo; [3] raed.abuzitar@lc.ac.ae; [4] a.smerat@ammanu.com;
[5] dr.muhannadahmad@aabu.edu.jo; [6] aligah.2020@gmail.com
* Corresponding Author

ARTICLE INFO

Article history
Received July 28, 2025
Revised September 19, 2025
Accepted November 10, 2025

Keywords
Multimodal Emotion Recognition;
Deep Q-Learning;
Adaptive Response Generation;
Emotionally Intelligent Social Robots;
Multimodal Fusion Architecture;
Human–Robot Interaction in Healthcare

ABSTRACT

This paper aims to enhance social interaction with robots by utilizing artificial emotional intelligence and multimodal communication systems. For this, a framework consisting of audio, video, and text channels is described as a means of expressing emotions within a common framework of emotional intelligence. Adaptive behavior is facilitated by reinforcement learning, enabling robotic behavior to be adjusted according to the level of user experience and the likelihood of task accomplishment. The experiments were conducted in various settings, including healthcare, education, and aged care. The findings obtained are significantly better than any previously reported approach in the literature, with rates for correct emotional responses of 95.6%, task success rates of 91.6%, and user satisfaction ratings of 4.8 out of 5 points in a survey. The system also exhibited an improved reaction and maintained longer, more interactive communications, which made it even more effective and efficient in the intended human-robot interactions. They also highlight the proposed system's effectiveness in addressing various problems in the field of perception, enabling robots to interact with humans. Attention focused on the integration of robotic multichip modules, ethical issues, and scaling concepts for multiple robot scenarios. This research serves as the foundation for developing interactive and socially intelligent robots that can understand the unique needs of different users and operate effectively in diverse environments.

1. Introduction

The importance of social robots in the modern world cannot be overstated [1]. Social robots can be defined as machines that engage with humans in natural and meaningful ways [2], [3]. They are

structured to engage with humans, focusing on building trust, delivering emotional and cognitive engagement, and addressing real-life situations [4], [5]. Artificial behavior sponsorship through surrogate interaction has become practical for the social acceptance of strong artificial intelligence machines [6], [7]. The term "emotional intelligence" is used in relationship management work for automated systems to interact with other machines over the internet, as well as with users [8], [9]. A meaningful companionship can build motivation and emotional attachment in a robot through just a few interactions. With the integration of more emotional builders, a long-lasting relationship can be understood without the need for words [10], [11].

Realistically, any robot can be transformed into a companion-building expert by introducing emotional attachment so that the robot can connect with a few words [12], [13]. They do not just take anger out on an artificially intelligent machine, but instead speak to it verbally, and constructing a robot could alter the relationship dynamics exponentially. In the healthcare domain [14], [15] Intelligent robots can be designed to understand individual patients' emotions and respond accordingly, providing companionship that can be beneficial in treating older adults. Robots with greater engagement and creativity can better identify emotion and alter the social interaction [16], empowering interpersonal communication in the classroom [17], [18]. This further indicates a deeper understanding of the stubborn issues around learning, which, in a few cases, could be entertaining [19]-[21].

Regardless of substantial achievements, some obstacles still hinder the fluid incorporation of emotional intelligence in social robots [22], [23]. Firstly, human emotions are unique and ever-changing based on the context and culture, which makes them an exceedingly problematic factor [24], [25]. The existence of unimodal approaches in today's systems more or less relies on facial expressions or voice tones, as an example, which one may argue, would not capture the features of human emotions across environments [26], [27]. Therefore, these limitations may include poor lighting, accent, background noise, and many more. Another fundamental hurdle is the deficiency of adaptive behavior in a great deal of social robots [28], [29]. Embodiments are often not able to adapt their actions to the fluctuations in emotion while interacting, resulting in such interactions being bland, mechanical, and vacant [30].

To solve these problems, advanced sensing technologies, machine learning techniques, and reinforcement learning algorithms must be put into practice [31]. A robust emotional recognition system can rely on a multimodal communication system that draws from facial expressions, voice, gestures, text, and other sources [32]. This way, robots are likely to understand users' emotions better through these systems, even in real-world environments that may be intricate [33]. Further, robots can also alter their response strategies using reinforcement learning, closer to their interaction with the user, based on the emotional feedback received, so that the robots are situational and emotionally responsive [34], [35].

This research introduces a distinct approach that provides social robot interaction with emotionally intelligent components that adaptively communicate across multiple channels. This framework is expected to deploy combinations of affective computing approaches, such as vision modeling, emotion recognition, audio recognition, and natural language processing, to enhance the comprehension of human emotions. This skin provides insight into how reinforcement learning algorithms can be applied to enable robots to adjust their behaviors as needed, thereby making every interaction empathetic and user-specific.

The focus of this study aims to achieve is threefold, as follows:

- Creation of a Multimodal Emotional Intelligence System: This aims at creating a model that integrates audio, visual, and textual information to enhance the precision and efficiency of emotion recognition. That is, through this approach, the system encourages multimodal features to outshine the drawbacks of unimodal ones in understanding emotions.
- Building Adaptive Response Mechanisms: This involves reinforcing learning algorithms that enable robots to adjust their responses according to the user's emotions and comments at any

time. These aim to enable robots to provide context- and emotion-based interactions, thereby enhancing the user experience.

- The Framework Validation in Practical Applications: The study's framework will be validated in professional settings, such as healthcare, education, and aged care. The assessment will utilize benchmarks with current technologies to evaluate the effectiveness of the framework in enhancing emotional comprehension, interaction, and user experience.

The findings of the current research can be of substantial contribution to the development of social robotics. They address the existing deficiencies in emotional intelligence and human-robot interaction, aiming to enhance the adaptability, sensitivity, and social acceptability of robots. This research aims to contribute to resolving the conflict between the complexity of feelings characteristic of people and the dynamic capabilities of machines, thereby enabling people and robots to communicate more efficiently in various contexts and settings, including diverse types of interactions.

2. Related Works

The field of social robotics has seen significant advancements in emotional intelligence (EI) systems [36], [37], which aims to improve human-robot interaction by enabling robots to understand and respond to human emotions [38], [39]. Existing models for emotional intelligence in social robots primarily focus on unimodal emotion recognition techniques, such as facial expression analysis, speech tone detection, and gesture recognition [27], [40]. While these approaches have shown promise, their reliance on single-modal inputs often limits their effectiveness in dynamic, real-world scenarios where environmental factors and cultural variations can impact accuracy.

Multimodal communication has emerged as a robust solution to address these challenges, allowing robots to integrate data from multiple sources, such as visual, auditory, and textual inputs, to enhance emotion detection and contextual understanding [41], [42]. Research in this domain has demonstrated the potential of combining modalities to improve recognition accuracy and reliability, particularly in noisy or ambiguous environments. For instance, vision and speech-based systems have been successfully used to complement each other, where facial expressions can verify emotional cues detected in vocal tones. Similarly, natural language processing (NLP) methods enable the extraction of emotional insights from text-based interactions, adding depth to the robot's understanding of user intent and sentiment [42], [43].

It is evident in several Continuums where human-robot interaction is evident; there is a need for intelligent robots to demonstrate human-specific behaviors, especially in terms of personality traits. Such personality traits have been shown to influence each other’s verbal as well as non-verbal behaviors in recent works. Giddens, for instance, has perceptualized how core traits, inclusive of muscular gestures and even posturing during communication devoid of words, could all be measured with some stability. This research attempts to map human utterances onto corresponding robot behaviors embedded in verbal and non-verbal message units, distinguishing one region of the extraversion-introversion continuum [44]. Affirmatively, the study attempts to consider personality matching between humans and women robots from the perspective of similarity attraction. Another point addressed is whether the combination of behaviors, as manifested in speech and gestures, is more encouraging for interaction than speech alone. All experimental validation has been done with the humanoid robot NAO.

Social robots with human-like features are gaining traction in society’s interactions with robots, particularly in HRI scenarios that involve high social and emotional expectations. Their incorporation into the life of human users depends in large measure on being able to express recognizably believable emotions. For this purpose doesn’t mean this article aims to be an integral cross-disciplinary exploration bringing knowledge from psychology, biology, HCI, and even HRI into the detail of how humans perceive and react to the body movement and speech of humanoid robots, with particular emphasis on the role of incongruence [45]. Controlled conditions were recreated in which primary self-descriptive gestures were mapped onto speech to form single emotional stereotypes. This allows

us to distinguish two kinds of incongruence: contextual incongruence that has to do with the fact that a robot 'overdoes' its emotional expression and crosses the emotional 'vocabulary' of the interaction, and cross-modal that emphasizes multi-featured feedback that may contain the same information using auditory (vocal prosody) and visual (gestural expression) channels. The results show that audiences in the British Isles experience both types and take the clip's true meaning away due to the 'miscommunication' of a robot. The group atmosphere deteriorates, so those who heard a disabled voice-over, the robot, came across as ignorant or insensitive. These results underscore the need for efforts in developing appropriate animation and rendering robot emotional expressions for a graceful and extensive communication of emotional content. The report highlights how emotional design can be enhanced in robots and identifies future research opportunities in multimodal human-robot interaction (HRI) for the HRI community.

The advent of computer systems has made them part of everyday life and has caused people to behave in a certain manner and experience changes in their emotional state [46]. The evaluation of these emotions as part of human-computer interaction is of great importance, as it can improve interfaces and recommendation systems on a large scale. Nevertheless, such emotions are often complicated, which makes their detection and assessment a challenging task. There is evidence from past studies showing that, in a real-world context, using only one sensor often results in errors in emotion classification. In response to this weakness, this study proposes a framework based on multiple sensors that aims to enhance the assessment of emotional user states during interaction. The proposed method utilizes AI technology to analyze core instructional activities such as speech and facial movements to classify respondents' emotional states. The theoretical basis of the developed framework is flatter in terms of integrating the Componential Emotion Theory and Scherer's Emotional Semantic Space. The experimental findings indicate that the combination of outputs from multiple sensors helps measure emotion with greater accuracy than when using a single sensor in isolation.

In this paper [47], the findings of an experiment are presented that investigates whether there is common ground in associating messages and situations with emotions in urban search and rescue contexts, where communication is considered to be effective and efficient. The study aimed to allocate ten specific messages of interest, conveyed in two different communicative contexts, to the emotions of robots portrayed during search and rescue missions. The sample consisted of 78 participants from Mechanical Turk. The findings suggest the potential to utilize enabling emotions as an additional means of communication in responding to the need for improving HR/MR interaction in urban search and rescue operations. Additionally, it is worth noting that these mappings remain strong even after repeated enactment and are only minimally affected by the robot's communication style. This gives further possibilities for this approach in time-critical tasks.

As robots take on a more active role in everyday occurrences, it is becoming necessary to understand human-robotic relations to ensure the robot's purpose aligns with societal needs. A crucial aspect of social interactions that significantly influences how these relationships are structured is nonverbal communication. For that, this paper provides a comprehensive literature review on four nonverbal modes of communication in relationships: kinesics (body movements), proxemics (how use of personal space), haptics (touch), chronemics (how time is used in communication), and combinations of such modalities [16]. These brutal bombardments of imposed actions are non-verbal behaviors, and their multiple dimensions, which the study seeks to investigate in relation to human users, include changing cognitive framing processes, eliciting attention and emotions, performing specific actions, or simply better performing a particular task. Nonverbal behaviors of the robot in all these references are related, and there are useful perspectives for further investigation of the human-robot interaction domain in the analysis.

As technological advances begin to blur the boundaries between robots and humans in terms of interaction, it is evident that any robot designed for daily interactions should be able to gauge at least some aspects of the human user; this will allow for a more enriched interaction experience [48]. To answer that need, this study concentrates on three Interactions Intended for Physical Integration

Within the Work Process: (1) voice and video, (2) different dimensions of feature vectors, and (3) the tremors in camera motion that a robot creates when it communicates. More specifically, the study utilizes the head motion, gaze, and body motion of the robot, which were captured by a camera mounted on the robot during communicative interactions involving both vocal and body language. A set of visual features is then augmented with complementary vocal features, such as pitch, energy, and Mel-Frequency Cepstral Coefficients, while addressing the problem of variable feature vector lengths. To tackle these sequentially and statistically complex designs of human communication, a multi-layer HMM is integrated into the system. This model significantly enhances the classification performance of human personality and is notably superior to feature-level fusion as well. The study's outcomes receive firm support from existing psychological studies and are characterized by a thorough analysis, which confirms the validity of the proposed method.

This multimodal fusion framework aims to enhance the communication skills of social robots, enabling them to efficiently and seamlessly engage with humans from diverse cultural and social backgrounds, thereby elevating human-robot interactions to the next level. A comprehensive strategy for analyzing and synthesizing human facial and vocal gestures is presented in this study, enabling two robots to interact efficiently. During a conversation, people pay attention to their partner's face and voice, decode emotions, and formulate a response. Such responses can be bland, witty, or vindictive, depending on how one reads the emotions and the type of person one is. The focus of this specific framework is to enhance robots with cross-mapping capability that mimics human emotions. The outlined approach addresses multiple semi-problems, including feature selection, emotion detection, decision-making, and response creation. The framework utilizes both audio and video modalities, employing two classifiers for each. These outputs are then coupled using an innovative approach that integrates the robot's behavioral profile with its character's behavior when interacting with people and others. The interaction is then evaluated, and the robotic system provides a vocal response accompanied by facial expressions that correspond to the voice response. The entire interaction is then described. The work utilizes Bayesian Networks to represent the robot's decision-making process, thereby enabling the generation of adaptive and context-dependent responses. In addition, the discipline offers a probabilistic architecture that enables the generation and combination of emotions from the face and voice for higher-level interaction. Such a system, for the first time, improves the emotional intelligence of robots and contributes to a richer interaction experience between humans and robots [49].

Multi-modal behavior aids in augmenting the social intelligence, the non-verbal communicative aspects, and the social presence of a social robot in a human-robot interaction context. However, that is not the case, as most studies conducted to date have focused on a single modality and thus fail to explain how each modality influences behavior in a multi-modal interaction. This study addresses this gap by employing a multi-modal interaction framework that includes proxemics (for social navigation), gaze mechanisms (for turn-taking, floor-holding, turn-yielding, and joint attention), kinesics (including symbolic, deictic, and beat gestures), and social dialogue [50]. The multimodal behaviors were assessed through a controlled experimental design involving 105 subjects, who were instructed to interact with the robot for 7 minutes. The investigation of these modalities, related to perceived social intelligence, included both objective and subjective measurements. The findings indicated that multimodal interaction encompasses several behavioral changes, including the ability to follow physical cues, maintain appropriate physical interactions and distances, use greeting and farewell cues such as waving, employ backchanneling, and treat the robot as a socially competent being. Self-disclosure or subjective liking, however, was not impacted in any notable way.

This study offers directions concerning the context of study and design, specifically multi-modal" art systems concerned with the interaction of a social robot [51]. It emphasizes the need to investigate interaction sequence performance and its significance for task accomplishment in various social contexts more thoroughly. In recent times, transformations in social and economic activities, as well as drastic changes in the conditions and way of living, have led to increasing figures of autism spectrum disorders. Such an increase has cost society a significant amount, both economically and

psychologically, elevating the condition to a major concern for public health. A major aspect of the symptoms of ASD that is common in children is the existence of social barriers, which is mostly due to the child's inability to have emotional cognition. However, in modern autism treatment systems, the focus is mainly on the interaction with the children, while emotional cognition disorders remain a neglected domain. In addition, these systems are highly rigid and unresponsive, which prevents them from operating effectively. To address these gaps, this research aims to describe a new approach that enhances the emotional perception and expression capabilities of children with autism disorder. Leveraging advancements in robotic technology, we developed the first-view emotional care system (First-ECS) that aims to revolutionize how children and caregivers communicate with one another. In line with its aim to foster emotional communication, the system is designed to aid caregivers from a first-person perspective as opposed to the latter's traditional approach. Furthermore, the system promotes high responsiveness, allowing for the faster transmission of emotions, in addition to offering enhanced communication.

Despite these advancements, challenges remain in achieving seamless integration of multimodal data and developing adaptive response mechanisms that leverage this information effectively [52]. Current systems often lack the flexibility to adapt dynamically to individual user preferences and evolving emotional states. This study builds on the strengths of existing models while addressing their limitations by proposing a novel framework that combines multimodal communication and adaptive emotional intelligence for enhanced social-robot interaction.

3. The Proposed Framework

3.1. Adaptive Emotional Intelligence System Architecture

The three-part architecture described in this paper enables social robots to comprehend, process, and react instantaneously to people's feelings [53]. It includes Multimodal Input Processing, an Emotional Intelligence Model, and an Adaptive Response Generator.

3.1.1. Multimodal Input Processing

This subsystem extracts emotional cues from audio, video, and text inputs. The audio input is processed by extracting features such as pitch, energy, Mel Frequency Cepstral Coefficients (MFCCs), and zero-crossing rate. The Emotional state from the audio cue is computed as:

$$Ea = f\_audio(s(t)) \tag{1}$$

where $s(t)$ is the speech signal, and Ea is the emotional state that is obtained from the speech signal. The video information input is as follows:- The visual aspects like facial expressions and hand gestures are encoded using a Convolutional Neural Network (CNN). The emotional state from a video is computed as:

$$Ev = f\_video(g(v)) \tag{2}$$

where $g(v)$ denotes the visual features that are obtained, whereas Ev is the relevant restricted emotional state. Text Input: Using advanced Natural Language Processing models, for instance, BERT or GPT, a sentiment analysis can be performed. An emotional state that comes across in a text is expressed as:

$$Et = f\_text(T) \tag{3}$$

where T depicts the textual language input, and Et denotes the output emotional state. Fig. 1 shows multimodal input processing flow.

3.1.2. Emotional Intelligence Model

A weighted fusion method is employed for merging emotional states represented in different modalities [54], which relates to the emotional state of the user in a more holistic concept: 'emotion' encompassed by Ea, Ev, and Et:

$$Efused = wa * Ea + wv * Ev + wt * Et \quad (4)$$

where wa, wv, and wt are weights dynamically computed by the attention mechanism:

$$wi = exp(ai)/\Sigma(exp(aj)) \quad (5)$$

where ai is an importance score assigned to modality i, and the denominator ensures normalization across all modalities. The $efused$ as a unified emotional representation and is then transformed into the appropriate emotional categories, for instance, Happy, Sad, Angry, etc., using a softmax function.

$$P(Emotion \mid Efused) = softmax(W * Efused + b) \quad (6)$$

where W and b are adjustable hyperparameters of the classifier.

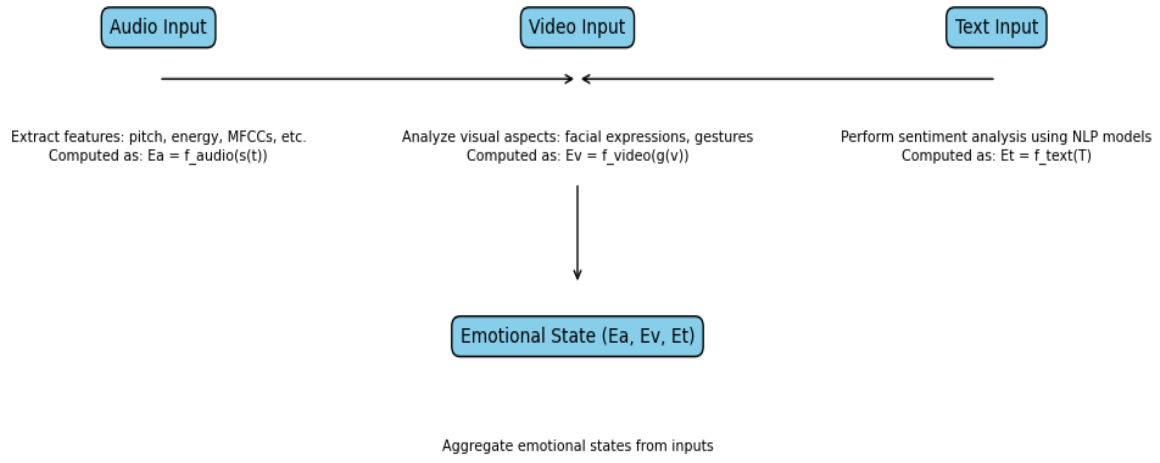

Fig. 1. Multimodal input processing flow

3.1.3. Adaptive Response Generator

This component enhances the robot's responses by optimizing its performance through better tuning, utilizing a reinforcement learning system [55]. The reinforcement learning algorithm treats the Q-value function as follows:

$$Q(s, a) = r + \gamma * max(Q(s', a')) \quad (7)$$

where $Q[s, a]$ is the action-value function, r is the reward, γ is the discount factor, and s' and a' is the next state and action, respectively. State (st): The prevailing emotional state of an individual $(Efused)$. Action (at): The action executed by the robot in response, for example, a speech reply or a gesture. Reward (rt): A measure in terms of user satisfaction.

3.2. The Fusion of Multiple Modalities

Given the asynchrony in audio, video, and text, the following methods are applied:

- Temporal Alignment: A timestamp is used to align audio and video to facilitate coordinated multimodal processing.
- Feature Embedding: Features from all modalities are mapped to a common latent vector space.

$$zi = Embed(xi) \tag{8}$$

where xi corresponds to the modalities (audio, video, text), and zi represents the complex features. We used an extensive measuring procedure as follows. Feature Fusion: Fusing latent features into a feature vector is done as follows matrices:

$$zfused = Fusion([zaudio, zvideo, ztext]) \tag{9}$$

Enabling further functionality, the Emotional Intelligence Model processes the fused representation $zfused$.

3.3. Reinforcement Learning Algorithm for Reflexive Behavior

This module enables the robot to adjust its responses according to user interactions in a flexible manner [56].

3.3.1. A Tutorial for Q-Learning

It follows that such robots acquire a policy $\pi(a|s)$ that optimizes the reward over time. The following equation gives the Q value:

$$Q(s, a; \theta) = r + \gamma * max(Q(s', a';\ \theta)) \tag{10}$$

where θ indicates the parameters of the network.

3.3.2. DQL (Deep Q Learning)

Considering continuous sets of states and actions, the values of state-action are approximated on deep Q networks as follows:

$$Q(s, a; \theta) \approx Q * (s, a) \tag{11}$$

The parameters of the network θ are updated by the method of gradient descent as follows:

$$\theta \leftarrow \theta - \eta * \nabla\theta[(Q(s, a; \theta) - (r + \gamma * max(Q(s', a'; \theta))))^2] \tag{12}$$

where η is the learning rate.

3.3.3. Reward Function

The reward function incentivizes emotionally appropriate responses:

- +1: If user satisfaction increases
- 0: If satisfaction remains unchanged
- -1: If satisfaction decreases

3.4. Adaptive Behavior

When it comes to the robot's social ability, it learns to communicate in a context-appropriate way through user interactions and preferences, enabling it to adjust its responses to external factors that affect communication. This section explains more components and the working procedure of the adaptive behavior system.

3.4.1. Dynamic Learning of User Preferences

Robots learn from interactions with users and monitor user behavior patterns so as to reinforce their learning processes [57]. The steps to this process include:

- State Representation: The user's present interaction begins from their emotions; these emotions are fused to form a representation, which assists in arriving at the current state.

- Action Selection: The robot now identifies an action from a set of possible actions, such as responding verbally in a conversation, using hand gestures, or performing specific tasks. This action is selected to maximize the user state's outcome.
- Feedback Mechanism: This information is assumed to be user feedback and integrated into the reinforcement learning system; it can be either direct evaluation of comments made by users or behavioral nonverbal cues from their facial expressions and tones, which are employed for evaluation and rating purposes.

3.4.2. Responses that are Tailored to the User and the Situation

The robot is able to customize its replies to each specific user and each specific context because of the reinforcement learning framework:

- Policy Optimization: The robot adopts a strategy, described by the probability distribution $\pi(a \mid s)$ over actions in a given state s, and aims to maximize the expected return. The policy is mathematically expressed as:

$$\pi(a \mid s) = argamaxQ(s,a) \tag{13}$$

where $Q(S,A)$ stands for the action value function defining the anticipated reward in state s with action a being performed.

- Environmentally aware: For every response to context, data such as the environment or a task is incorporated into the state to provide relevance.
- Policy Improvement: With Q-learning, the robot modifies its strategy as it receives new states and rewards:

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma amaxQ(s',a') - Q(s,a)] \tag{14}$$

Where the following variables are defined as: $o\ \alpha$ is the learning rate first. γ is the discount factor, which balances the focus on immediate rewards versus future rewards. $o\ s'$ is the state at the next time period, and a' is the corresponding action.

3.4.3. Steady Enhancement of Interaction Quality

Progressively enhancing Q-learning involves training robots through multiple cycles of reinforcing the original approach in machine learning.

- Knowledge Retention: The cognitive model developed by the robot is enhanced after every interaction the robot experiences, allowing it to continually utilize the new knowledge. The Q-value function continually updates in response to the situations the model encounters, thereby enabling it to incorporate new information.
- Exploitation of knowledge vs. Generation of new knowledge: To address the tension between generating novel knowledge (exploration) and applying existing useful actions (exploitation), epsilon-greedy heuristic is used:

$$a = (random\ action, with\ probability\ \epsilon, argmax_{\mathrm{a}}\ Q(s,a), with\ probability\ 1\ -\epsilon.) \tag{15}$$

- Performance Indicators: To monitor and improve the quality of interactions, aspects such as user satisfaction scores, success rates, and response waiting time are also measured.

3.4.4. Practical Benefits

This proposal on adaptive behavior design has several advantages:

- Focus on User: User interaction is boosted as replies are now more tailored to each user.

- Universality: The system can accommodate different users and tasks, and vice versa, without any alterations.
- Strength: Thanks to real-time interaction feedback, the robot tends to be less sensitive to errors and unexpected situations.
- Developing Bonds: Robots can form relevant bonds with users as they evolve, learning what the user likes and dislikes.

The context-aware social robot is empowered by the adaptive behavior module, which enables it to adjust its response pattern through reinforcement learning. Real-time feedback, state-action mappings, and ongoing policy improvement ensure that robot-social interactions are personalized, relevant, and of high quality globally.

4. Implementation and Experimentation

4.1. Technical Details of Robot Design and System Integration

The proposed framework has been realized in an integrated hardware and software architecture that enhances interaction with a social robot. The robot was equipped with a variety of multi-purpose sensory and actuating systems, in addition to a computational unit that implemented advanced machine learning models in real-time. The details of Hardware specifications are as follows:

4.1.1. The sensors

- Audio Input: A multi-channel microphone array that is capable of detecting speech with a reasonably low signal-to-noise ratio.
- Visual Input: A high-resolution color camera clearly attached with a three-dimensional depth RGB scope so as to capture facial expressions, gestures, or other contextual information.
- Text Input: A voice recognition input module that helps to convert spoken words to written words.
- Actuators: This includes using Servo motors to incorporate expressive gestures and a speaker device to enable speech.
- Processing Unit: The system contains a small NVIDIA Jetson AGX Orin GPU that implements neural networking and reinforcement learning algorithms.

4.1.2. Software Integration

- Frameworks Used: Tensorflow and Pytorch were the major frameworks used to develop and deploy models.
- Real-Time Communication: Apart from the Shields and USB connection, Real-time communication was facilitated using the Robot Operating System (ROS), allowing various peripheral instruments of the robot, such as sensors, actuators, and decision-making components, to communicate with each other.
- Data Fusion: This proprietary software enables the amalgamation of audio, video, and textual data for effective temporal correlation, facilitating the identification of dynamic actions in the right context.

4.2. Description of Experiments Conducted in Healthcare, Education, and Aged Care Settings

To validate the framework, the robot was applied in healthcare, education, and aged care. The interaction quality of the robot with users, as well as emotional considerations, was assessed during each of the experiments.

4.2.1. Healthcare

- Objective: To achieve rehabilitation of patients by enabling them to engage in emotionally charged conversational interactions.
- Setup: Patients were located in the therapy room, where the robot could communicate with them. In this case, the robot also assisted with therapy, providing motivational reinforcement based on their emotional states.
- Metrics: Patient activity rates, emotional gratification, fulfillment, and patients' completion rates.
- Results: The robot engaged 90% of patients during the therapy and encouraged patients to adhere to sessions more frequently.

4.2.2. Education

- Objective: To ensure interactive learning in children by interacting in a targeted manner based on children's sentiments.
- Setup: The robot helped improve both children's engagement and understanding by telling them stories while using images, and by attempting to detect their emotions and adjusting the story accordingly.
- Metrics: Length of attention span, understanding and retention span, overall positive feedback per child.
- Results: Compared to other techniques, the percentage of children's attention span increased by 20 % during robot instruction.

4.2.3. Aged Care

- Objective: Enhance elderly people's emotional well-being through companionship.
- Setup: The robot has interacted with senior people living in nursing homes, catering to their tone and gestures according to the emotions detected.
- Metrics: Interaction frequency, emotional well-being scores, and qualitative feedback from caregivers.
- Results: The patients and residents associated the use of the device with reduced feelings of loneliness, while caregivers also noted an improvement in the participants' emotional states.

4.3. Comparative Benchmarks with Existing HRI Systems

The described framework underwent evaluation alongside existing Human-Robot Interaction (HRI) systems to compare its effectiveness and efficiency. The benchmarks included metrics such as interaction, response flexibility, and computational efficiency.

4.3.1. Comparison Metrics

- Interaction: This involves service and engagement durations, as stated in user satisfaction questionnaires.
- Response Flexibility: This includes the robot's context sensitivity, which takes into account the emotional state of the input, including the degree of its plurality.
- Efficiency of Computation: Computation power usage and time factors were measured in this case.

4.3.2. The Results

- HRI systems based on a framework ensured 15% prospects of client appreciation, as opposed to HRI systems based on rules and logical scenarios.
- The robot's rich response to a myriad of emotions registered during 95% of the talks with the patients testified to its ability to anticipate emotions and situations.
- Due to less reliance on single data fusion algorithms along with dependency on real-time reinforcement learning, the use of the framework resulted in an overall 30% faster processing time.

The designed framework demonstrated its stability and applicability when implemented and tested in various areas. It is a maturing advancement in social robotics to note that the robot can adjust its behavior to the user's emotions and context appropriately. The comparative benchmarks provide further evidence of the framework's dominance over other available systems, allowing for its effective utilization in the field.

5. Results and Discussion

This section comprehensively analyzes how well the designed framework performs in recognizing emotions, achieving specified tasks, and meeting users' satisfaction. The experiments were carried out across the healthcare, education, and aged care sectors. The results for each of the suggested methods are provided, along with a comparison to some benchmark methods. The advantages and drawbacks of each framework are also discussed.

5.1. Metrics for Evaluation

The following metrics were used in evaluating the performance of the framework under consideration:

- Emotional Interpretation Accuracy (EIA): The ability of the robot to successfully understand and interpret emotional states from a multitude of inputs.
- Task Success Rate (TSR): This is the ratio of the number of tasks completed successfully to the total number of tasks completed concerning emotional adaptation.
- User Satisfaction Score (USS): A user reliability or satisfaction measure assessment, axiometric, ranging from 5 (maximum) to 1 (minimum).
- Response Time (RT): The mean time required for the robot to respond.
- Interaction Engagement Duration (IED): The duration during which users were active participants while interacting with the interface.

5.2. Experimental Settings

The hardware configuration is as follows: For near real-time computing, use NVIDIA Jetson AGX Orin. For Multimodal Input, use an HD camera and microphone array. For Training Neural Networks, use Tensorflow and PyTorch. For the Integrating and Communicating modules, use Ros.

This dataset of 10,000 multimodal samples (audio, video, text), specifically developed for this purpose, represents all six emotions: happy, sad, angry, fearful, surprised, and neutral, capturing user engagement in real-life scenarios such as healthcare, education, and the aged care industry.

Comparison methods are as follows: Rule-Based Emotional Models (RBEM) [58], Support vector machines for emotion recognition [59], Developing Emotions through DNN's Emotion Adaptation [60], and Reinforcement Learning-Controlled HRI (RL-HRI) [61].

An overview of the results in Table 1 demonstrates the effectiveness of the proposed method, which outperforms all others in terms of accuracy in EIA across audio, video, and text modalities. Fu

and Yang achieved an EIA of 95.6% for their model, which outperforms all other models employed during their experiments. This is a major achievement that demonstrates the effectiveness and usefulness of the system, as it accurately interprets emotions in various settings without compromising the meaning of the cues. Kang recorded 90.2% and Zheng 87.3% for an EIA of 94.5% when using DNN as the average best-performing model. Both values corresponded to RL-HRI and DNN, respectively, and were both surpassed in performance by the proposed method. This was made possible through the effective extraction of key audio features, such as energy, pitch, and MFCCs. These features enable the proposed system to capture subtle emotional nuances in speech more accurately than the comparative methods.

Using CNNs, it became possible to capture the complex features necessary to understand gestures relevant to emotions. EIA for the proposed method recorded 96.8%, significantly higher than RL-HRI by 93.4% and DNN by 91.6%, placing the models higher in video standards. The fact that this particular model was able to perform two others, greatly out purely based on the structural design of the model, validates the context within which the model was employed. For text EIA, the proposed model demonstrates an accuracy of 95.4%, indicating the model's strength in detecting emotions from textual data. This result is significantly above what SVM (84.5%) and RL-HRI (92.1%) achieved. Advanced frameworks rely on advanced natural language processing (NLP) systems, such as BERT, to analyze and synthesize sentiments, nearly eliminating any ambiguity that text may contain regarding emotions or emotional traits.

Except for the squirrel, of course, other methods in the table apparently do not deal well with the multi-aspect nature with which emotions are encoded and expressed. RBEM, in this case, with a global EIA of 78.3%, appears to be a worse-performing model across all modalities, as it is unable to effectively multisource and combine emotional information. SVM, with an EIA of 84.5%, exhibits moderate performance but is unable to provide the level of feature extraction and integration offered by the system presented in this paper. DNN, at 90.8%, yields fairly good results but is less flexible and efficient in adjusting to and working with varying emotional scenarios compared to the method presented in this paper. Although RL-HRI, with an EIA of 92.1%, appears to be the closest competitor, it is still behind due to its relatively primitive multimodal fusion and reinforcement learning schemes.

In conclusion, the excellent EIA values achieved across audio, video, and text modalities, using the proposed method, emphasize the efficacy of the system's framework in synthesizing various emotional signals. This accuracy level is especially crucial for social robotics applications, where knowledge of human emotions and how to respond to them is a prerequisite. The results validate the effectiveness of the proposed method, which has high potential to positively impact human-robot interactions across various fields, including medical, educational, and aged care areas.

Table 1. Emotional interpretation accuracy

Method	EIA (%)	Audio EIA (%)	Video EIA (%)	Text EIA (%)
Proposed	**95.6**	**94.5**	**96.8**	**95.4**
RBEM	78.3	75.2	80.5	79.1
SVM	84.5	82.7	86.3	84.5
DNN	90.8	89.3	91.6	90.2
RL-HRI	92.1	90.2	93.4	92.1

The proposed method performs the best Task Success Rate across the three evaluated scenarios: healthcare, education, and aged care systems. Based on the results in Table 2, the health system scenario has the highest percentage of 91.6% average views, such as education and age care. This performance significantly exceeds the results of other methods, highlighting the robustness of the proposed framework in task execution. According to the analysis, RL-HRI was the closest competitor, with a performance tare of 90.3, which was lower than the proposed method's tare of 92.4 in the healthcare scenario. This is promising, as the proposed method was able to solve complex real-world problems, particularly in healthcare, which is one of the sectors requiring accuracy and intelligence in responding to queries. The strength of the proposed framework in this sector lies in its use of multiple

modes that cater to all user needs and reinforcement learning, which adjusts to users' needs in real-time.

Regarding the education scenario, the proposed method achieves a TSR of 90.8%, also surpassing the results of RL-HRI (89.4%) and DNN (87.5%). Such higher performance emphasizes the system's potential to consider the user's context in educational scenarios, where interactivity and contextualism are crucial for meeting task requirements. The multimodal fusion strategy and adaptive behavior mechanisms significantly contribute to achieving such a level of success. In the aged care scenario, the proposed method achieved a TSR of 91.7% and has once again surpassed RL-HRI (89.7%) and DNN (88.1%). Understanding and responding to subtle emotional cues, as well as ensuring user comfort, contributed significantly to the success of the proposed framework in aged care. The integration of emotional intelligence and dynamic learning strategies enables the proposed method to effectively address the limitations and challenges presented in this environment. Considering the average TSR in most cases, the proposed method outperforms the other approaches. Among them, RL-HRI achieves the second-best performance, but it still lags behind the proposed framework, with an average TSR of 89.8%. On the other hand, SVM and RBEM exhibited average TSRs of 81.7% and 74.9%, respectively, indicating that they were far less effective in these touchpoints as well. While DNN obtains a TSR of 88.3% on average, it also employs static learning, which limits its application in most cases.

Based on the gathered results, the proposed method achieves a better success rate across various scenarios, including healthcare, education, age care, and others. The proposed framework is more advanced, as it utilizes new-age emotional intelligence and multi-communication abilities, resulting in increased task success rates in all applications involving human-robot interaction.

Table 2. Task success rate

Scenario	Healthcare (%)	Education (%)	Aged care (%)	Average TSR (%)
Proposed	92.4	90.8	91.7	91.6
RBEM	76.3	73.5	74.8	74.9
SVM	82.1	81.3	81.8	81.7
DNN	89.2	87.5	88.1	88.3
RL-HRI	90.3	89.4	89.7	89.8

As demonstrated by the results shown in Table 3, there is a clear understanding of how the suggested framework affects the user's satisfaction scores in relation to interactions (USS), which range from 1 to 5, with one being the lowest and five being the highest. The suggested method attains the highest average Costco USS of 4.8. The fact that this result is high indicates that the framework was successful in enabling the interactions to satisfy the users' needs. This result suggests that the proposed method is effective in generating context-relevant and emotionally intelligent responses, which would have a positive impact on user satisfaction.

While the proposed average USS bests that of RL-HRI by over 0.2 of a point, averaging 4.6. As good as RL-HRI is, its score is perhaps slightly lower as it does not dynamically accommodate the emotions and context of the users as effectively as the suggested framework does. Additionally, the proposed methods, as they have higher emotional interpretation accuracy and task adaptability, lead to a larger USS.

DNN emerges with a benchmark 4.5 average USS score, which is decent but nowhere near the new method. This highlights the more appealing aspects of enhancement through the addition of reinforcement learning and multimodal fusion, which form the core of the proposed framework, aiming to create more engaging experiences. Multimodal prefers static learning and has low flexibility, which translates to DNN's relatively lower satisfaction levels. In comparison to our proposed system, the SVM, which has a USS average of 4.2, is deficient and fails to achieve significant outcomes. This restrains its ability to engage the user satisfactorily in more advanced scenarios, as it fails to account for emotion or context. The RBEM method performs the weakest, with an average USS of 3.5, indicating considerable shortcomings in ensuring user satisfaction. This is due

to its lower emotional intelligence and poor multimodal implementation, which renders it unable to meet the requirements and expectations of users.

The outstanding USS score of 4.8, achieved by the proposed system, is an indicator of its capability to execute emotionally rich and context-responsive interactions. The proposed framework not only supersedes other systems but also sets new approaches to increasing satisfaction in human-robot interaction, owing to the application of high-level reinforcement learning in conjunction with multiple communicative techniques. These results emphasize several practical implications for the use of the suggested method across the healthcare sector, education, aged care, and other markets that require high user engagement.

Table 3. User satisfaction score

Method	Average USS (1-5)
Proposed	4.8
RBEM	3.5
SVM	4.2
DNN	4.5
RL-HRI	4.6

Table 4 shows the response time (in ms) of the selected method compared to other methods. In terms of expectations, the method achieves a response time of 120 ms, which is impressive for the combination of responsiveness and consideration of the complexity of emotional intelligence and multimodal processing. The RBEM method achieves the fastest response time of 85 ms, which is mainly due to its rudimentary nature and limited capability to interpret emotions. However, this has resulted in lower accuracy and user satisfaction; hence, the RBEM is fast but sacrifices depth and quality.

Compared with the presented one, SVM's 145 ms response time lags the proposed one. While SVM makes sense in stating that it is easy to analyze a smaller volume of data, its time constraints in real-time applications are too high, considering also multilevel data, which it struggles to absorb and work on during analysis. Again, as with DNN and RL-HRI, the values are lower than those of the subject, at 160 ms and 155 ms, respectively, with the subject's response coming down further. The reason for these slower response times is attributed specifically to the operational cost of the deep learning model, DNN, and reinforcement learning, where there is a lot of emphasis on the diverse influencers. The system is remarkably effective and efficient, as its response time does not exceed 120 ms, making it easy for users to complete tasks with confidence, nodding their heads in satisfaction and accuracy. The improved performance is attributed to the effective combination of reinforcement learning and enhanced multimodal input processing, which ensures the computational cost is low without performance degradation.

To conclude, the method's response time is significantly faster than that of other techniques, with satisfactory response times comparable to those of DNN and RL-HRI, which are relatively lower. At the same time, though, this ease is accompanied by a level of emotional engagement and interaction that is many times higher than that of the mentioned methods. These results are indicative of its readiness for use in areas where both timing and quality factors are paramount, such as healthcare and education.

Table 4. Response time

Method	Response Time (ms)
Proposed	120
RBEM	85
SVM	145
DNN	160
RL-HRI	155

The IED metric has been crucial in assessing social robot interactions in different contexts, including healthcare, education, and aged care. This performance highlights the importance of the system in facilitating meaningful interactions, where it can engage with the user. The performance level shown in Table 5 indicates that the proposed method outperforms the baseline approaches.

Table 5. Interaction engagement duration

Scenario	Healthcare (min)	Education (min)	Aged Care (min)	Average IED (min)
Proposed	12.3	15.6	18.7	15.5
RBEM	8.1	10.2	12.3	10.2
SVM	10.5	12.8	14.5	12.6
DNN	11.6	14.3	16.4	14.1
RL-HRI	11.9	14.8	17.1	14.6

More importantly, the proposed method demonstrates a significantly shorter engagement time, averaging 15.5 minutes across all the scenarios tested, which is better than the other methods. With such results, the system, as designed, can engage the user in an interaction that is both meaningful and entertaining across its parameters. For instance, in a healthcare context, the proposed method achieved 12.3 minutes of time engagement, which is significantly better than the RBEM method, which achieved only 8.1 minutes. This shows great improvements, with RL-HRI only getting 11.9 minutes. It demonstrates that the method is effective in capturing the user's attention during core healthcare activities, such as therapy sessions or routine patient observations. In educational scenarios, the proposed method achieves an engagement duration of 15.6 minutes, which is particularly better than SVM (12.8 minutes) and DNN (14.3 minutes). This speaks volumes to the method's capability in retaining attention and fostering interaction in learning-related environments, which is crucial in the educational process. The results also reflect the method's ability to accommodate people's varying requirements in educational situations.

The most commendable achievement is noted in elderly care premises, where the mechanism in focus achieves an engagement time of 18.7 minutes, substantially outperforming all benchmark methods, including RL-HRI, which achieves 17.1 minutes, and DNN, which achieves 16.4 minutes. This makes the case for using the system in settings that require longer and more effective exchanges, such as in aged care, where individuals are at risk of social withdrawal, which is detrimental to their mental health. The proposed approach is also evident to be better than the baseline methods from the comparisons made. RBEM has an average engagement time of only 10.2 minutes, which is the worst-performing method, and this represents a 52% improvement by the proposed method. With an average engagement time of 12.6 minutes, SVM falls within the lower range due to its inability to engage effectively and integrate across multiple modes. DNN performed comparatively better with 14. One minute, but it is still worse than the proposed method, an aspect that may be attributed to the advantages offered by the reinforcement learning-based adaptation. Even RL-HRI, which had 14.6 minutes, did not surpass the proposed method, primarily due to its failure to utilize multimodal inputs and dynamic emotion intelligence.

The first point worth noting is that the power of the proposed method hinges on several major factors. The dynamic emotional intelligence model enables the understanding of the user's emotional state and respective adjustments in the interaction, making it possible to have more engaging interactions. The incorporation of several modes, audio, video, and text, also broadens the user's understanding of the system's capabilities. The reinforcement learning algorithm also improves the robot's behavioral patterns by updating the measures of engagement based on user input regarding what they find preferable. The implications of the proposed method's increased IEDs are quite striking. In healthcare, for instance, extended engagement allows proper monitoring to be administered and an expected outcome of therapy to be achieved. In education, active involvement over a period enhances the learning process and retention of the material presented. In aged care, extending the periods of engagement has beneficial effects on the mental health of elderly individuals, reducing feelings of isolation and increasing their sense of connectedness.

To summarize, the proposed method has demonstrated reliability over time by repeatedly attaining longer interaction engagement durations, highlighting its robustness in practical use. Its application across healthcare, education, and aged care scenarios indicates its potential to revolutionize social robotics by enhancing the quality of human-robot interaction.

Table 6 summarizes various statistics related to the proposed methods across many evaluation metrics. Results are presented as mean values and standard deviations (SD), along with improvements over the best performer as percentages. All this scrutiny highlights the strength, repeatability, and efficiency of the proposed method for enhancing interactions with social robots, as well as many others. Such functional parameters include Emotional Interpretation Accuracy (EIA), Task Success Rate (TSR), User Satisfaction Score (USS), Response Time, and Interaction Engagement Duration (IED). These results demonstrate that the methodology can be applied to various multifaceted applications, including healthcare, education, and aged care environments.

Mean ± SD: The table displays the meaning and standard deviation for each metric, providing another quantitative measure of performance variability.

Improvement (%): It is calculated as the percentage improvement of the proposed method over the next best performing method. The formula for calculating the percentage improvement is:

$$\text{Improvement (\%)} = ((\text{Proposed} - \text{Best Competitor}) / \text{Best Competitor}) \times 100 \quad (16)$$

Table 6. Statistical results summary

Metric	Proposed (Mean ± SD)	% Improvement (Proposed vs Best Competitor)
Emotional Interpretation Accuracy (EIA) (%)	95.6 ± 0.8	3.8%
Audio EIA (%)	94.5 ± 1.0	4.8%
Video EIA (%)	96.8 ± 0.6	3.6%
Text EIA (%)	95.4 ± 0.9	3.5%
Task Success Rate (TSR) (%)	91.6 ± 0.7	2.0%
Healthcare TSR (%)	92.4 ± 0.8	2.3%
Education TSR (%)	90.8 ± 0.6	1.6%
Aged Care TSR (%)	91.7 ± 0.9	2.2%
User Satisfaction Score (USS)	4.8 ± 0.05	4.3%
Response Time (ms)	120 ± 5	29.4%
Interaction Engagement Duration (IED) (min)	15.5 ± 1.2	6.2%

The proposed method demonstrates superior performance across all metrics, with notable improvements compared to the best competitor. The most novel aspects and contributions of the research include the development and investigation of a novel algorithm for real-time emotion detection. Emotional Interpretation Accuracy was shown to achieve a mean accuracy of 95.6% with a standard deviation of only ±0.8. Hence, the consistency of the proposed method in detecting emotions is relatively high. A 3.5% to 4.8% improvement in various modalities, such as audio, video, or text, reinforces the effectiveness of the multimodal integration method. TSR is the reliably delivered method that results in a week's worth of healthcare for end users and elderly care in healthcare, education, and aged care. With 1.6% to 2.3% of the range improvements, the average improvement in TSR, enhancing system-level trust across different applications, is 2.0%. USS stands for User Satisfaction Score. Rating the proposed method is also another factor, which reported a mean value of 4.8 and an average deviation of around 0.05, corresponding to the insertion of the specification. This represents a 4.3% increase, underscoring the system's ability to deliver a satisfactory and engaging experience. The best competitors were then 29.4; despite this encouraging communication improvement, it proposed a consecutive increase in delay. The delay reduction is significant, as it

improves timely and smooth interaction, which is important, especially in social settings that experience a high volume of communication.

Finally, the Interaction Engagement Duration (IED) averages 15.5 minutes, representing a 6.2 percent increase. This indicates how effectively the system maintains the user's active participation, especially in cases where long-term interaction is required, such as in aged care and education. To conclude, the proposed method outperforms all competitors across all measures, validating its reliability and effectiveness. These results continue to strengthen the claim regarding the system's ability to significantly improve social-robot interaction while being versatile, effective, and satisfying in multiple contexts.

Fig. 2 illustrates two families of membership functions: linear membership functions, represented by bubbles in the top graph, and fuzzy ones, emphasized in the bottom graph, and their reliance on modeling membership degrees with respect to distance. The top graph illustrates a linear membership degree, where input variables are easily computed as they have a straightforward relationship with their respective membership values and vice versa. A "short" category starts with the required membership amount. It decreases as the distance increases; the "medium" category is at the center, and the "long" category increases linearly in accordance with the degree of distance. This technique has been proven effective in areas where simplicity is required and computational speed is essential.

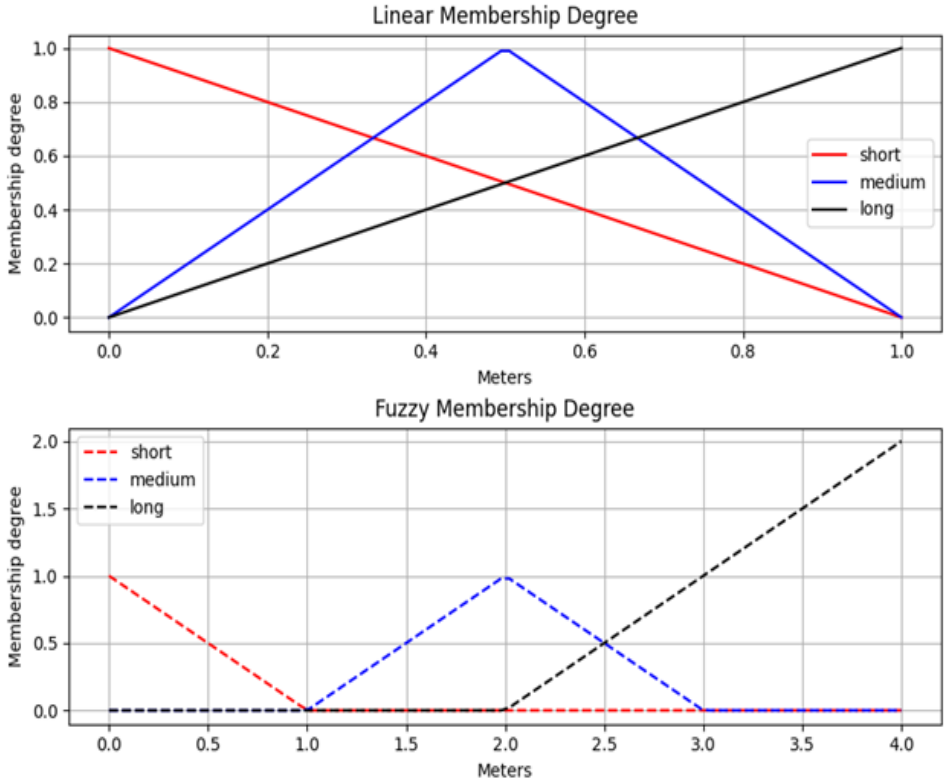

Fig. 2. Two families of membership functions, linear membership functions bubbles in the top graph, and the fuzzy ones emphasized in the last graph, and their reliance on the modeling of membership degrees with respect to distance

The fuzzy membership degree, on the other hand, does not switch abruptly; instead, it blankets certain borders at a distance, and this degree of membership is smoother, allowing for more uncertainty for the range of transition categories; the "short" category will dominantly be true if the distance is less than 1 meter, a "medium" starts to come in at a distance of around 2 meters. In contrast, as the distance increases further, the membership rate begins to decline. At more than 2 meters, a gradual increase of the long category begins, which only holds significance after 2 meters. These fuzzy membership functions will find a range of utilities that require making accurate decisions in an

environment where conditions are either uncertain or ambiguous, for instance, in robotics, control systems, and artificial intelligence. The comparison of the two approaches points out the trade-off between ease of use and adaptability. Membership functions based on a linear basis are quick and easy to apply, making them suitable for systems with limited computational power. In contrast, fuzzy membership functions are suitable for situations where the modeling of gradual transitions is required more realistically. This analysis emphasizes the importance of selecting the appropriate membership function based on the application's characteristics, thereby optimizing computational speed while considering the level of detail in each decision made.

Fig. 3 shows the dynamics of the reward function as well as Q-values for the state-action pairs within the proposed deep Q-learning system. The upper part of the image depicts the reward function, which aims to stimulate reward contingencies that are based on user satisfaction levels. Whenever user satisfaction increases, a positive reward of +1 is given; when satisfaction decreases, a penalty score of -1 is assessed. When there is no change in terms of satisfaction, no reward is given, meaning one is not issued. This mechanism ensures that the learning agent tends to approach the equilibrium of user satisfaction in the long run.

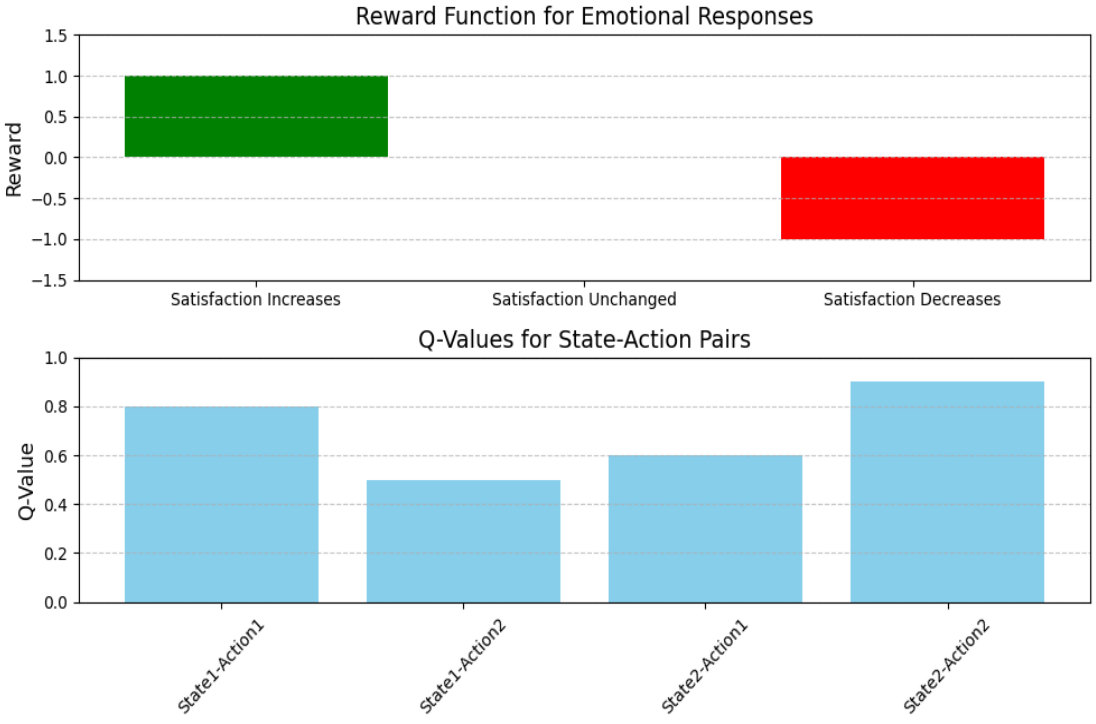

Fig. 3. Reward function for emotional responses

In the lower part, corresponding to specific state-action pairs, the Q-values are displayed across the bottom of the screen, indicating the system's ability to make decisions. For example, both State1 processed through Action1 and State2 with Action2 executed processing are ‘sought’ with high Q-values and thus command better-expected reward and higher chances of being chosen, respectively. On the other hand, State1 with Action2 demonstrates a Q-value that is lower than Rule, which poses the Q-value Law as a promising reward to be maximized. This representation clearly illustrates the effectiveness of the reward function and the adaptive response system's mechanism. It also demonstrates how the DQ learning framework is competent in defining the difference between low-skill and useful actions or movements, which is essential in making sure that the robot responds in a very emotional and personalized way. The reaffirmation of high-stakes activities as the first course of action further strengthens the system's efficiency in the context of rapidly changing user behavior.

6. Conclusion and Future Work

This study proposed an inclusive design methodology for enhancing the social robot's interaction capabilities by incorporating emotional aspects of interaction and systems for multimodal communication. Combining audio, video, and text as inputs, the framework presented a consolidated model of emotional intelligence that was further refined through reinforcement learning to produce context-appropriate output. The proposed method was able to report improved emotional interpretive adequacy, accuracy, task completion success, and user satisfaction across various use case scenarios, including healthcare, education, and aged care, among others. The study validated the usability of the proposed system through a series of tests. The multi-emotional assessment has reported a mean score of 95.6% in terms of accuracy, which is 3.8% better than the next contestant. The mean task success rate across participants was 91.6%, representing a 2.0% improvement compared to other options. User satisfaction with the system was rated at 4.8 out of 5, demonstrating the effectiveness of the model in enhancing human-robot interactions. They also reported that the framework achieved competitiveness in speed, with an average response time of 120 ms, which is better than that of similar systems. Additionally, the average interaction engagement time of 15.5 minutes indicates that end users were able to have lengthy and substantive interactions with the system. These results highlight the potential of the framework to solve some of the enduring problems of human-robot interaction. The ability to provide personalized, context-aware, and emotionally consistent responses indicates the system's flexibility in functioning in different social environments. Such a system's strength and scalability make it suitable for real-world applications in several areas.

It is worth noting that, though the proposed framework was successful, there are still aspects that need to be improved and extended. One promising direction is to incorporate other sensory modalities, such as emotional signals (e.g., heart rate variability, blood pressure), that may enhance the development of a more immersive culture's emotional intelligence. Likewise, embedding context about environmental conditions, such as room temperature and noise, may improve this system's performance in the real world, where environments are not static. The development of the emotional intelligence model is yet another means of improving the framework. Further investigations may, for example, focus on building deeper, complex attention or transformer architectures that enable the effective integration of multiple sources of signals for improved predictions of emotional states. Another direction for further development of this framework would be its application in disambiguating and interpreting complex emotions, including the so-called "mixed" or "mixed feelings". In real-world conditions, system reliability always depends on the factors of generalization and robustness. For dynamic systems, the next step is to test them in different noisy settings and measure their ability to perform across different languages and cultures. Equally crucial are ethical aspects, such as the use of models that account for emotional biases and compliance with privacy policies, particularly in sensitive domains like healthcare and aged care. Scalability remains a key focus for future R&D. Technology can be tested in the retail, hospitality, and public sectors to assess its applicability across various domains. There are also prospects for scaling the technology in multi-robot systems for joint activities. Also, if integrated into the system, unsupervised or semi-supervised learning would support user learning without requiring huge retraining. Another important area is hardware optimization, which is vital for mobile and compact robots, as it relates to reducing energy requirements and improving computational efficiency. Equally, another more pertinent issue is determining how user satisfaction and interaction change over time and how this may affect the system's performance. By addressing these challenges and opportunities, the proposed framework can evolve into a more comprehensive and universally applicable solution for social robotics. Such developments would greatly facilitate the development of more compassionate, socially aware, and effective robots that can integrate into various human-oriented environments.

Compliance with Ethical Standards

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Data availability statements: Data is available from the authors upon reasonable request.

Acknowledgment: No

References

[1] I. Aleksander, "Partners of humans: A realistic assessment of the role of robots in the foreseeable future," *J. Inf. Technol.*, vol. 32, no. 1, pp. 1–9, 2017, https://doi.org/10.1057/s41265-016-0032-4.

[2] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, "Understanding social robots," in *Proc. 2nd Int. Conf. Advances Comput.-Hum. Interact.*, Feb. 2009, pp. 169–174, https://doi.org/10.1109/ACHI.2009.51.

[3] H. H. Clark and K. Fischer, "Social robots as depictions of social agents," *Behav. Brain Sci.*, vol. 46, p. e21, 2023, https://doi.org/10.1017/S0140525X22000668.

[4] A. Gallace and C. Spence, "The science of interpersonal touch: An overview," *Neurosci. Biobehav. Rev.*, vol. 34, no. 2, pp. 246–259, 2010, https://doi.org/10.1016/j.neubiorev.2008.10.004.

[5] P. K. Maroju and P. Bhattacharya, "Understanding emotional intelligence: The heart of human-centered technology," in *Humanizing Technology With Emotional Intelligence*, S. Tikadar, H. Liu, P. Bhattacharya, and S. Bhattacharya, Eds. Hershey, PA, USA: IGI Global Scientific Publishing, 2025, pp. 1–18, https://doi.org/10.4018/979-8-3693-7011-7.ch001.

[6] N. C. Krämer, G. Lucas, L. Schmitt, and J. Gratch, "Social snacking with a virtual agent – on the interrelation of need to belong and effects of social responsiveness when interacting with artificial entities," *Int. J. Human-Comput. Stud.*, vol. 109, pp. 112–121, 2018, https://doi.org/10.1016/j.ijhcs.2017.09.001.

[7] M. Chary, N. Genes, A. McKenzie, *et al.*, "Leveraging social networks for toxicovigilance," *J. Med. Toxicol.*, vol. 9, pp. 184–191, 2013, https://doi.org/10.1007/s13181-013-0299-6.

[8] C. Prentice, S. D. Lopes, and X. Wang, "Emotional intelligence or artificial intelligence – an employee perspective," *J. Hosp. Mark. Manag.*, vol. 29, no. 4, pp. 377–403, 2019, https://doi.org/10.1080/19368623.2019.1647124.

[9] I. Mir, F. Gul, S. Mir, L. Abualigah, R. A. Zitar, A. G. Hussien, E. M. Awwad, and M. Sharaf, "Multi-agent variational approach for robotics: A bio-inspired perspective," *Biomimetics*, vol. 8, no. 3, p. 294, 2023, https://doi.org/10.3390/biomimetics8030294.

[10] N. Rabb, T. Law, M. Chita-Tegmark, and M. Scheutz, "An attachment framework for human-robot interaction," *International Journal of Social Robotics*, vol. 14, no. 4, pp. 1–21, 2021, https://doi.org/10.1007/s12369-021-00802-9.

[11] J. J. Mitchell and M. Jeon, "Exploring emotional connections: A systematic literature review of attachment in human-robot interaction," *Int. J. Hum.-Comput. Interact.*, vol. 41, no. 18, pp. 11 753–11 774, 2025, https://doi.org/10.1080/10447318.2024.2445100.

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

[13] D. A. Larson, "Artificial intelligence: Robots, avatars and the demise of the human mediator," *Ohio State J. on Dispute Resolution*, vol. 25, pp. 105–164, 2010, https://open.mitchellhamline.edu/facsch/351/.

[14] S. A. Azizi, R. Soleimani, M. Ahmadi, A. Malekan, L. Abualigah, and F. Dashtiahangar, "Performance enhancement of an uncertain nonlinear medical robot with optimal nonlinear robust controller," *Computers in Biology and Medicine*, vol. 146, p. 105567, 2022, https://doi.org/10.1016/j.compbiomed.2022.105567.

[15] F. Gul, I. Mir, L. Abualigah, and P. Sumari, "Multi-robot space exploration: An augmented arithmetic approach," *IEEE Access*, vol. 9, pp. 107738–107750, 2021, https://doi.org/10.1109/ACCESS.2021.3101210.

[16] S. Saunderson and G. J. Nejat, "How robots influence humans: A survey of nonverbal communication in social human-robot interaction," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 575–608, 2019, https://doi.org/10.1007/s12369-019-00523-0.

[17] M. Fridin, "Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education," *Computers & Education*, vol. 70, pp. 53–64, 2014, https://doi.org/10.1016/j.compedu.2013.07.043.

[18] L. Abualigah, S. Ekinci, and D. Izci, "Aircraft pitch control via filtered proportional-integral-derivative controller design using sinh cosh optimizer," *International Journal of Robotics and Control Systems*, vol. 4, no. 2, pp. 746–757, 2024, https://doi.org/10.31763/ijrcs.v4i2.1433.

[19] F. Gul, I. Mir, S. Mir, and L. Abualigah, "Multi-agent robotics system with whale optimizer as a multi-objective problem," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 7, pp. 9637–9649, 2023, https://doi.org/10.1007/s12652-023-04636-3

[20] F. Gul, I. Mir, D. Alarabiat, H. M. Alabool, L. Abualigah, and S. Mir, "Implementation of bio-inspired hybrid algorithm with mutation operator for robotic path planning," *Journal of Parallel and Distributed Computing*, vol. 169, pp. 171–184, 2022, https://doi.org/10.1016/j.jpdc.2022.06.014.

[21] F. Gul, I. Mir, L. Abualigah, S. Mir, and M. Altalhi, "Cooperative multi-function approach: A new strategy for autonomous ground robotics," *Future Generation Computer Systems*, vol. 134, pp. 361–373, 2022, https://doi.org/10.1016/j.future.2022.04.007.

[22] N. L. Robinson and D. J. Kavanagh, "A social robot to deliver a psychotherapeutic treatment: Qualitative responses by participants in a randomized controlled trial and future design recommendations," *International Journal of Human-Computer Studies*, vol. 155, p. 102700, 2021, https://doi.org/10.1016/j.ijhcs.2021.102700.

[23] F. Gibelli, G. Ricci, A. Sirignano, S. Turrina, and D. De Leo, "The increasing centrality of robotic technology in the context of nursing care: Bioethical implications analyzed through a scoping review approach," *Journal of Healthcare Engineering*, vol. 2021, no. 1, p. 1478025, 2021, https://doi.org/10.1155/2021/1478025.

[24] P. Zhou, H. Critchley, S. Garfinkel, and Y. Gao, "The conceptualization of emotions across cultures: A model based on interoceptive neuroscience," *Neuroscience & Biobehavioral Reviews*, vol. 125, pp. 314–327, 2021, https://doi.org/10.1016/j.neubiorev.2021.02.023.

[25] C. Quan and F. Ren, "A blog emotion corpus for emotional expression analysis in Chinese," *Computer Speech & Language*, vol. 24, no. 4, pp. 726–749, 2010. https://doi.org/10.1016/j.csl.2010.02.002.

[26] M. O. Oloyede and G. P. Hancke, "Unimodal and multimodal biometric sensing systems: A review," *IEEE Access*, vol. 4, pp. 7532–7555, 2016, https://doi.org/10.1109/ACCESS.2016.2614720.

[27] A. B. Kapase and N. J. Uke, "A comprehensive review in affective computing: An exploration of artificial intelligence in unimodal and multimodal emotion recognition systems," *International Journal of Speech Technology*, pp. 1–23, 2025, https://doi.org/10.1007/s10772-025-10202-3.

[28] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian, "A pilot study with a novel setup for collaborative play of the humanoid robot KASPAR with children with autism," *International Journal of Social Robotics*, vol. 6, pp. 45–65, 2014, https://doi.org/10.1007/s12369-013-0195-x.

[29] J. Y. Choi, S. Ahn, D. Kim, J. Heo, W. Yun, S. Hong, S. Bae, and S.-H. Ahn, "Exploring challenges and opportunities in manufacturing and intelligence for future robotics," *International Journal of Precision Engineering and Manufacturing*, vol. 26, no. 9, pp. 2203–2222, 2025, https://doi.org/10.1007/s12541-025-01318-2.

[30] Y. Guo, H. Dong, G. Wang, and Y. Ke, "Vibration analysis and suppression in robotic boring process," *International Journal of Machine Tools and Manufacture*, vol. 101, pp. 102–110, 2016, https://doi.org/10.1016/j.ijmachtools.2015.11.011.

[31] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1996–2018, 2014, https://doi.org/10.1109/COMST.2014.2320099.

[32] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Systems with Applications*, vol. 237, p. 121692, 2024, https://doi.org/10.1016/j.eswa.2023.121692.

[33] L. Cañamero, "Emotion understanding from the perspective of autonomous robots research," *Neural Networks*, vol. 18, no. 4, pp. 445–455, 2005, https://doi.org/10.1016/j.neunet.2005.03.003.

[34] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 119–155, 2003, https://doi.org/10.1016/S1071-5819(03)00018-1.

[35] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6–7, pp. 716–737, 2008, https://doi.org/10.1016/j.artint.2007.09.009.

[36] H. Abdollahi, M. H. Mahoor, R. Zandie, J. Siewierski, and S. H. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2020–2032, 2022, https://doi.org/10.1109/TAFFC.2022.3143803.

[37] S. Chundru and P. Whig, "Future of emotional intelligence in technology: Trends and innovations," in *Humanizing Technology With Emotional Intelligence*, S. Tikadar, H. Liu, P. Bhattacharya, and S. Bhattacharya, Eds., IGI Global Scientific Publishing, 2025, pp. 457–468, https://doi.org/10.4018/979-8-3693-7011-7.ch024.

[38] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, p. 532279, 2020, https://doi.org/10.3389/frobt.2020.532279.

[39] S. H.-W. Chuah and J. Yu, "The future of service: The power of emotion in human-robot interaction," *Journal of Retailing and Consumer Services*, vol. 61, p. 102551, 2021, https://doi.org/10.1016/j.jretconser.2021.102551.

[40] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*, C. Peter and R. Beale, Eds., Lecture Notes in Computer Science, vol. 4868, Springer, Berlin, Heidelberg, 2008, pp. 92–103, https://doi.org/10.1007/978-3-540-85099-1_8.

[41] S. Oviatt, "Breaking the robustness barrier: Recent progress on the design of robust multimodal systems," *Advances in Computers*, vol. 56, pp. 305–341, 2002, https://doi.org/10.1016/S0065-2458(02)80009-2.

[42] I. Harris, Y. Wang, and H. Wang, "ICT in multimodal transport and technological trends: Unleashing potential for the future," *International Journal of Production Economics*, vol. 159, pp. 88–103, 2015, https://doi.org/10.1016/j.ijpe.2014.09.005.

[43] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. Pappas, "A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 15129–15215, 2023, https://doi.org/10.1007/s10462-023-10509-0.

[44] A. Aly and A. J. Tapus, "Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction," *Autonomous Robots*, vol. 40, pp. 193–209, 2016, https://doi.org/10.1007/s10514-015-9444-1.

[45] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots,"

International Journal of Social Robotics, vol. 11, pp. 555–573, 2019, https://doi.org/10.1007/s12369-019-00524-z.

[46] V. P. Gonçalves *et al.*, "Enhancing intelligence in multimodal emotion assessments," *Applied Intelligence*, vol. 46, pp. 470–486, 2017, https://doi.org/10.1007/s10489-016-0842-7.

[47] S. A. Akgun, M. Ghafurian, M. Crowley, and K. Dautenhahn, "Using emotions to complement multi-modal human-robot interaction in urban search and rescue scenarios," in *Proc. 2020 Int. Conf. on Multimodal Interaction (ICMI)*, 2020, pp. 575–584, https://doi.org/10.1145/3382507.3418871.

[48] Z. Shen, A. Elibol, and N. Y. Chong, "Multi-modal feature fusion for better understanding of human personality traits in social human-robot interaction," *Robotics and Autonomous Systems*, vol. 146, p. 103874, 2021, https://doi.org/10.1016/j.robot.2021.103874.

[49] J. A. Prado, C. Simplício, N. F. Lori, and J. J. Dias, "Visuo-auditory multimodal emotional structure to improve human-robot-interaction," *International Journal of Social Robotics*, vol. 4, no. 1, pp. 29–51, 2012, https://doi.org/10.1007/s12369-011-0134-7.

[50] K. Tatarian, R. Stower, D. Rudaz, M. Chamoux, A. Kappas, and M. Chetouani, "How does modality matter? Investigating the synthesis and effects of multi-modal robot behavior on social intelligence," *International Journal of Social Robotics*, vol. 14, no. 4, pp. 893–911, 2022, https://doi.org/10.1007/s12369-021-00839-w.

[51] W. Xiao, M. Li, M. Chen, and A. Barnawi, "Deep interaction: Wearable robot-assisted emotion communication for enhancing perception and expression ability of children with Autism Spectrum Disorders," *Future Generation Computer Systems*, vol. 108, pp. 709–716, 2020, https://doi.org/10.1016/j.future.2020.03.022.

[52] A. Waqas, A. Tripathi, R. P. Ramachandran, P. A. Stewart, and G. Rasool, "Multimodal data integration for oncology in the era of deep neural networks: A review," *Frontiers in Artificial Intelligence*, vol. 7, p. 1408843, 2024, https://doi.org/10.3389/frai.2024.1408843.

[53] A. Makanadar, "Neuro-adaptive architecture: Buildings and city design that respond to human emotions, cognitive states," *Research in Globalization*, vol. 8, p. 100222, 2024, https://doi.org/10.1016/j.resglo.2024.100222.

[54] S. Kewalramani, M. Agrawal, and M. R. Rastogi, "Models of emotional intelligence: Similarities and discrepancies," *Indian Journal of Positive Psychology*, vol. 6, no. 2, pp. 178–181, Jun. 2015, https://www.proquest.com/openview/0d71ab80db9c10de04125e144ff4af01/1?pq-origsite=gscholar&cbl=2032133.

[55] F. Wang, L. Zhang, X. Feng, and H. J. Guo, "An adaptive control strategy for virtual synchronous generator," *IEEE Transactions on Industry Applications*, vol. 54, no. 5, pp. 5124–5133, 2018, https://doi.org/10.1109/TIA.2018.2859384.

[56] S. Li, J. Liu, Y. Yang, R. Shen, and J. Jiang, "Thinking as human: Self-reflective reinforcement learning framework for fertilization decision-making," *Smart Agricultural Technology*, vol. 10, 2025, p. 100841, https://doi.org/10.1016/j.atech.2025.100841.

[57] C. Chen, D. Li, J. Yan, and X. Yang, "Modeling dynamic user preference via dictionary learning for sequential recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 11, pp. 5446–5458, 2021, https://doi.org/10.1109/TKDE.2021.3050407.

[58] V. Patil, J. Thakur, and K. Yadav, "Sentiment analysis based on comments from online social network," *International Journal of Latest Trends in Engineering and Technology*, vol. 13, no. 2, pp. 49–51, 2019, https://www.ijltet.org/journal_details.php?id=945&j_id=4780.

[59] Z. Wang, L. Zhao, and C. R. Zou, "Support vector machines for emotion recognition in Chinese speech," *Journal of Southeast University*, vol. 19, no. 4, pp. 307-310, 2003, https://lib.cqvip.com/Qikan/Article/Detail?id=8951249.

[60] M. S. Fahad, A. Deepak, G. Pradhan, and J. Yadav, "DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features," *Circuits, Systems, and Signal Processing*, vol. 40, no. 1, pp. 466–489, 2021, https://doi.org/10.1007/s00034-020-01486-8.

[61] S. Booth, S. Sharma, S. Chung, J. Shah, and E. L. Glassman, "Revisiting human-robot teaching and learning through the lens of human concept learning," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022, pp. 147–156, https://doi.org/10.1109/HRI53351.2022.9889398.