TiagoCare: Emotion-Aware Interaction for Assistive Robotics

Elective in AI
HRI + RBC Report

Maria Emilia Russo*[1], Martina Doku*[1], and Giuseppina Iannotti*[1]

[1]Sapienza University of Rome
{russo.1966203, doku.1938629, iannotti.1938436}@studenti.uniroma1.it

*All authors contributed equally to this work.

1 Introduction

1.1 Context and Motivation

Healthcare environments such as hospitals, pediatric wards, and elderly care facilities often face a common challenge: ensuring continuous, personalized support for patients who may be experiencing psychological distress, emotional discomfort, or cognitive overload. Factors like medical uncertainty, isolation, and environmental stressors can exacerbate emotional vulnerability, negatively impacting both recovery and overall well-being. At the same time, resource constraints frequently make it difficult to provide the level of personalized human interaction required for optimal care. This creates a compelling opportunity for assistive technology, especially socially-aware robots, to augment existing workflows and offer emotional support to patients. However, deploying robots in healthcare remains a complex task. Barriers include patient trust, the challenge of integrating into clinical routines, and the technical difficulty of enabling robots to understand and appropriately respond to human emotional states. Human behavior is subtle, dynamic, and context-dependent; interpreting it correctly is challenging even for trained professionals, let alone autonomous systems. This project responds to that challenge by developing a modular robotic framework that enables intuitive and affect-aware interaction in emotionally sensitive environments. Built around the TIAGo robot platform, the proposed system integrates real-time perception, context detection, and symbolic reasoning to support adaptive behavior tailored to individual patient needs.

1.2 Objectives

The primary objective of this project is to demonstrate the capacity of a socially-aware robotic system to perceive human emotional cues and respond in a way that is both contextually appropriate and emotionally supportive. The system focuses on real-time multimodal emotion recognition, behavior modulation based on symbolic reasoning, and continuous personalization through user profiles and historical trends. In practice, this means enabling the robot to detect affective states such as sadness or anxiety, interpret the environmental and social context, and perform empathic actions like offering reassurance or maintaining an appropriate distance. These behaviors are evaluated in structured HRI simulations, testing the robot's ability to adapt based on diagnosis, context, and modality-specific cues. Ultimately, the project aims to contribute a robust and extensible architecture for affective human-robot interaction that can be deployed in real-world care settings to alleviate emotional distress and improve patient experience.

1.3 Summary of the Results

Throughout the development and testing of the TiagoCare system, we achieved a functional integration of perception, context modeling, and behavior generation. The emotion recognition modules reached average accuracies of 62% (face), 97% (audio), and 69% (pose) on standard benchmarks. Context detection using YOLOv8 and face identification achieved high reliability in typical indoor scenes. For symbolic reasoning, a lightweight knowledge graph embedding model trained on synthetic triples showed strong internal consistency (training loss = 0.032), enabling the robot to infer emotional states even when sensor data was inconclusive. In simulated interactions with 7 synthetic user profiles, the robot demonstrated personalized behavior based on diagnosis, environment, and emotional input. User study feedback (via Likert scores and qualitative comments) confirmed improved comfort and perceived empathy when personalized strategies were used. These results validate the system's ability to adapt socially and emotionally in real time, even under uncertainty, and demonstrate the value of combining symbolic reasoning with multimodal perception in assistive robotics.

2 Related Works

In recent years, the development of emotionally intelligent robotic systems has gained significant traction, particularly in the context of healthcare and assistive services. Hospitals and care facilities are emotionally demanding environments, where patients often experience stress, anxiety, or disorientation. Recognizing and mitigating such emotional states can lead to better health outcomes, improved communication, and reduced workload for caregivers. Social robots equipped with the ability to perceive and respond to human emotions—through cues such as facial expressions, vocal tone, and body posture—can offer crucial support in these scenarios.

More broadly, emotion-aware robotics holds promise in everyday applications, from home assistance for the elderly to educational and therapeutic interventions. These systems can foster more natural and trustworthy human–robot interaction, especially when designed to recognize user stress and adapt their behavior accordingly.

The TiagoCare system was conceived to explore this potential by combining multimodal emotion recognition, symbolic reasoning over contextual knowledge, and real-time interaction on the TIAGo robot platform. Our approach was grounded in two core thematic areas of the course: Human–Robot Interaction (HRI) and Robot Benchmarking and Competitions (RBC). The HRI module inspired the design of interaction strategies based on multimodal perception, empathy, and context awareness. The RBC module guided the development of the internal architecture, affective memory models, and symbolic world modeling using knowledge graphs.

Our goal is to evaluate whether combining real-time multimodal emotion recognition with symbolic scene and affect graphs enables a robot to exhibit more adaptive, empathetic behavior in emotionally charged environments. Drawing on the insights from [3] and [5], we were particularly interested in testing how the fusion of facial, vocal, and postural signals—processed in parallel—could improve the detection of stress and emotional ambiguity compared to unimodal pipelines. These works emphasized the benefits of using deep learning architectures for accurate and robust affective state estimation, especially when supported by redundancy across modalities.

At the same time, our system seeks to move beyond perception by embedding emotion-related knowledge into structured, symbolic graphs that evolve over time. Inspired by the context-aware modeling proposed in [19] and [13], we adopted a hybrid architecture that links sensor-derived affective cues to semantic information about the environment and the user's interaction history. This allowed us to investigate whether knowledge graph–based reasoning (e.g., using TransE) could generalize affective predictions across different users and contexts.

While our approach confirms established findings on the value of multimodal emotion recognition—both in terms of technical performance and user engagement—it adds novel contributions to the field. In particular, our work differs from prior implementations by tightly integrating affective inference with sym-

bolic reasoning and deploying the entire system on a resource-constrained service robot like TIAGo. Moreover, we propose a stress-centered design that prioritizes emotional de-escalation and personalized interaction, which remains an underexplored dimension in current affective HRI literature.
The following subsections present the most relevant academic and course references that informed these choices, organized according to the two main parts of the course: HRI and RBC.

2.1 HRI

The conceptual design of TiagoCare was significantly influenced by the Human-Robot Interaction course lectures [1], particularly Lecture 4, which introduced the principles and challenges of multimodal interaction. Multimodal systems, as emphasized in the lecture, provide a richer and more robust user experience by combining different communication channels (e.g., voice, gesture, visual cues). The course highlighted benefits such as increased usability, redundancy in communication, and improved accessibility, while also discussing the inherent complexities in integrating multiple technologies and modalities. A key theoretical framework presented in the lecture is the **MODIM approach** (Multi-modal interaction manager), developed during the COACHES project. MODIM formalizes interactions using templates composed of atomic communication and robot actions, abstracted from specific modalities. It enables defining interactions that are portable, maintainable, and customizable for different user profiles (e.g., age, language, occupation). This approach informed our architectural decisions by encouraging template-based modeling of robot–user dialogues and inspired our future intent to extend TiagoCare's reasoning layer with formalized interaction structures and multiple synchronized modalities. The idea of mapping abstract interaction actions to concrete multimodal outputs, was particularly useful in structuring our communication modules. Although TiagoCare does not yet integrate MODIM directly, the principle of parallel multimodal communication execution (e.g., speech output with GUI support and visual feedback) is mirrored in our implementation. Building on these conceptual foundations, our system was further shaped by the academic literature. In particular, we identified two major areas of prior work that inspired both the structure and functionality of TiagoCare:

- **Multimodal Emotion and Stress Recognition**, which informed the signal processing, fusion strategies, and CNN-based classification modules we adopted to detect user stress and emotional states.
- **Human-Robot Interaction in Assistive Contexts**, which provided architectural and behavioral models for designing empathetic, context-aware robotic assistants capable of adapting to user needs in healthcare and domestic environments.

The following subsections outline the most relevant contributions in each of these domains and clarify how they informed our design choices and implementation strategy.

Multimodal Emotion and Stress Recognition The core capability of TiagoCare—detecting stress and emotional ambiguity via multimodal signals—is inspired by recent advances in emotion recognition through deep learning. In particular, [3] provides a comprehensive overview of facial, audio, and physiological modalities, emphasizing the limitations of unimodal systems and the superior performance of multimodal fusion. A more practical affective HRI system is detailed in [5], where three Dynamic Bayesian Networks (DBNs) independently process facial and speech features before being fused to infer the user's emotional state. This model informed the early stages of our design, particularly in its layered architecture, facial ROI processing, and decision-level multimodal fusion. Our system adopts a similar temporal reasoning strategy, replacing DBNs with modern deep architectures to improve robustness and generalizability. Moreover, the modular system presented in [8] directly inspired our integration strategy. The authors use CNNs for visual and audio emotion classification and Transformer-based models for textual signals, with user-specific fine-tuning and large language model (LLM) synthesis. We adapted the concept of context-aware multimodal fusion and personalized feedback, leveraging lightweight CNNs (e.g., MultiNetV2) and a modular architecture compatible with ROS-based deployment on TIAGo. The MEC-HRI system introduced in [14] reinforces the importance of multimodal pipelines for recognizing and responding to human emotion. The paper validates the use of speech, facial, and gesture recognition with multiple layers of fusion (feature-level and decision-level), which served as a reference for building our flexible input processing modules.

Human-Robot Interaction in Assistive Contexts TiagoCare's focus on adaptive interaction in healthcare is rooted in the literature on assistive HRI. [14] describes an intelligent interface for elderly support that fuses audio and visual signals to enhance command understanding, which influenced our approach to handle speech and posture under noisy, real-world conditions. A more comprehensive system is proposed in [12], integrating spatial and transactional intelligence to enable context-aware services in domestic environments. The robot manages a knowledge base of past interactions and environmental constraints. We were directly influenced by this architecture when designing our adaptive reasoning layer and contextual emotion tracking system, which allow Tiago to modulate behavior based on emotional state, location, and prior user history. Finally, the work [6] highlights how social robots can deliver multimodal, emotionally sensitive assistance to elderly users—principles that align with the mission of TiagoCare.

2.2 RBC

The RBC course and its lecture materials [2] offered conceptual and technical guidance on cognitive architectures, reasoning systems, and affective computing—all of which informed the internal logic and memory models of TiagoCare.

In particular, Lecture 9 on Knowledge Graphs provided the foundation for representing contextual and emotional knowledge as directed graphs of entities and relationships. This inspired our implementation of a scene and affect graph, enabling the robot to reason over time, track user state, and adapt its actions accordingly.

Affective Computing and Context-Aware Robotics Affective computing principles underlie the reasoning and personalization mechanisms of TiagoCare. [19] extends the EMOTIC model with time-series analysis, enabling the robot to adapt to evolving emotional dynamics. This directly informed our implementation of temporal smoothing and emotion history tracking, critical for understanding user stress patterns and de-escalating interactions. [13] proposes a Bi-LSTM-based framework to capture valence–arousal transitions across conversational segments. Their use of context loss and emotional anchoring inspired us to build a long-term representation of user affect across interaction windows. From a personalization perspective, [22] describes a two-stage system for adapting to user feedback through pseudo-labeling and fine-tuning. While we do not support full retraining, our system leverages emotion-informed behavior personalization to adjust Tiago's speech and gesture parameters. [25] introduces LRP-based explainability for CNNs in emotion prediction. While not yet implemented, the idea of transparent emotional reasoning will be considered in future iterations of the system.

TIAGo in Healthcare and Social Robotics The technical feasibility of our work was supported by [17], where TIAGo's microphone array is used for real-time audio emotion detection. Additionally, [23] benchmarks CNNs on TIAGo, using EfficientNetV2 for fast inference with GUI integration. Our system builds on this by focusing on stress-specific emotion recognition and real-time behavioral adaptation.

Metrics and Evaluation Strategy In the literature, functional evaluation of emotion recognition modules typically relies on classification metrics such as *accuracy*, *F1-score*, and *confusion matrices*. For facial emotion recognition, the FER2013 dataset [24] is a common benchmark, often used in conjunction with CNN-based architectures. In our case, we adhered to this standard by training and validating our face module on FER2013 using accuracy as the main metric. For speech-based emotion classification, our module leverages a pretrained Wav2Vec2 model (`r-f/wav2vec-english-speech-emotion-recognition`) available on the Hugging Face Model Hub. This model enables real-time inference through a streamlined classification pipeline, making it suitable for interactive HRI scenarios. Its performance is reported directly by the model authors on unseen evaluation data drawn from the TESS [18], RAVDESS [15], and SAVEE [11] datasets. Body posture and gesture recognition are often assessed through precision, recall, and accuracy using datasets like BEAST [9] or MOBOT-6a; in TiagoCare, we used BEAST with KNN [10] classifiers and accuracy-based scoring. Multimodal emotion recognition systems are generally evaluated using

decision-level fusion metrics and ablation studies to determine modality contribution. We used a majority voting strategy combined with context-based fallback logic, and measured robustness through task success rather than only classifier performance. From a task-based evaluation perspective, standard HRI benchmarks focus on interaction fluency, user satisfaction, and goal completion. Rather than limiting our assessment to offline model metrics, we adopted a holistic strategy: a task was considered successful when the robot could correctly interpret the user's emotional state, generate semantically valid knowledge graph triples, and perform an appropriate, context-aware response. This approach differs from the state-of-the-art in two ways. First, it shifts the focus from isolated classifier performance to *end-to-end interaction quality* in realistic scenarios. Second, it incorporates symbolic reasoning and affect history into the evaluation loop, enabling qualitative insights such as user-perceived empathy and fallback effectiveness. While our qualitative feedback is informal, it reflects an emerging trend in HRI evaluation that emphasizes emotional alignment and human-centric outcomes over raw numerical scores. In summary, TiagoCare aligns with established RBC evaluation practices for individual perception modules, while advancing a more integrated and socially grounded testing procedure that accounts for the complexities of real-world assistive interaction.

Originality vs. Confirmation Our system builds upon prior findings that highlight the benefits of multimodal fusion and deep learning–based emotion recognition. However, its originality lies in the integration of symbolic reasoning, affective memory, and real-time robotic behavior into a unified pipeline. As detailed above, the use of RDF-style graphs, emotion history tracking, and graph-based inference allows TiagoCare to operate beyond simple classification—enabling context-aware, explainable, and adaptive interaction. In doing so, we contribute not just an improvement in technical accuracy, but a shift toward holistic and human-centered evaluation in affective HRI.

3 Integrated Solution

TiagoCare integrates a suite of perception, reasoning, and actuation modules to deliver socially adaptive behavior in real-time. The architecture enables the robot to perceive emotional and contextual cues, reason about user states, and respond appropriately through speech, movement, and gestures. This section provides a detailed overview of the information flow between modules, how memory and user models are maintained, and how the robot's behavior is shaped by social context. Section 3.1 examines human–robot interaction (HRI), and Section 3.2 evaluates performance based on the Robot Behavior Capability (RBC) framework.

Functional Architecture of the Solution

The TiagoCare system is composed of modular subsystems implemented as independent ROS nodes, each communicating via standardized topics and messages. The core architecture (Fig. 1) is structured around three main stages: **social signal processing**, **context detection**, and **social reasoning**, which collectively drive real-time, emotionally intelligent behavior.
Emotion perception is handled by the *Social Signal Processing* module, which analyzes facial expressions, vocal prosody, and body posture. Detailed implementation of these components is provided in Section 4. Emotional predictions from all modalities are passed to the *Multimodal Fusion* module, which unifies them into a dominant emotional state through majority voting.
In parallel, the *Context Detection* module gathers environmental and situational information using object detection and face recognition. This includes symbolic scene descriptions like `(person_1 sits on chair_3)` or `(person_1 is in waiting room)`, which are encoded as knowledge triples and embedded using PyKEEN's TransE algorithm.
Combined affective and contextual information is forwarded to the *Social Reasoning* module, which selects an appropriate behavior policy based on symbolic rules and user-specific profiles stored in memory. Resulting decisions are passed to the *Social Signal Generation* module, which handles speech, gestures, and proximity control. This modular pipeline allows the robot to adapt its behavior in real-time, demonstrating user-sensitive interaction strategies. A coordinating simulation node manages synchronization and execution across all modules.

3.1 HRI

The TiagoCare system is designed to detect and respond to diverse human social signals in real time, including facial expressions, vocal tone, body posture, and gestures. **Facial expressions** are captured via camera and classified into emotional categories such as happiness, sadness, and fear. **Vocal cues** are analyzed to extract prosodic features linked to internal states like stress or anxiety. **Body and hand pose** is used to identify postural indicators of emotional engagement. In parallel, the system gathers contextual information such as user

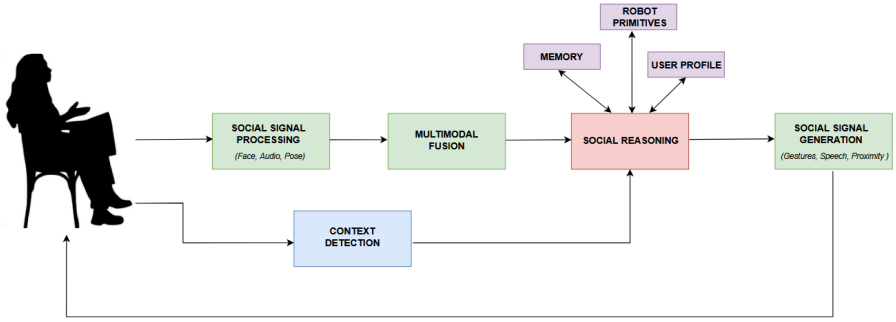

Figure 1: Functional Architecture of TiagoCare

activity and room layout using object detection and scene composition.
Human–robot interaction is both *passive and active*: the robot continuously observes the user and also prompts interaction through speech and gestures. Communication includes synthesized speech (via `espeak` or `pyttsx3`), predefined gestures (via `play_motion`), and symbolic control of proximity (`"close"`, `"medium"`, `"distant"`).
To support personalization, TiagoCare maintains structured user profiles that store identity, diagnosis, preferences, and emotional history. These are combined with real-time perception data and contextual understanding to form a symbolic world model. This model, updated dynamically during interaction, enables the robot to reason about the user's state and respond appropriately.
The *Social Reasoning* module selects behavior strategies based on current emotion, environment, and user preferences. If a preferred action is stored in the profile, it overrides generic decisions. Final behaviors are translated into concrete actions and dispatched to execution modules.
This architecture supports socially intelligent behavior such as empathy, adaptation, and situational awareness. For example, an anxious user may be met with a calm, supportive gesture, while another preferring low interaction will be respected with minimal engagement. Additional scenarios are detailed in Section 5.1.2.

3.2 RBC

The Robot Behavior Capability (RBC) analysis of the TiagoCare system focuses on evaluating both the technical functionality of individual modules and the system's ability to deliver socially appropriate, context-aware behavior. While experimental results are reported in the Results section, this subsection outlines the evaluation framework, module responsibilities, and how symbolic reasoning and world modeling contribute to adaptive behavior generation.
Functionally, each perception and reasoning module is benchmarked under controlled conditions using open-source datasets and simulated real-time sensor

input. Core components include facial emotion recognition, audio-based emotion classification, body and hand pose analysis, environmental perception, and symbolic reasoning based on knowledge graphs. These modules are assessed individually based on criteria such as accuracy, classification consistency, response latency, and robustness across varying environmental conditions.

At the task level, the system is validated using a structured simulation with seven synthetic user profiles, each reflecting a clinical diagnosis, emotional tendencies, and preferred robot behavior. These profiles enable controlled testing of behavior strategies across diverse scenarios, such as anxiety in clinical settings or low social tolerance in shared spaces. Fallback strategies are considered valid when resulting behaviors remain socially appropriate.

A central component enabling such adaptability is the system's symbolic world model. This model consists of dynamic knowledge graphs built from real-time perception inputs, represented as semantic triples (e.g., `(user_1, is_in, waiting_room)` or `(user_1, has_emotion, sad)`). These representations are embedded into a low-dimensional vector space using the TransE algorithm via PyKEEN, allowing the system to reason analogically and infer missing relationships or emotional states.

Knowledge is acquired through sensor fusion, user profiles, and environmental analysis. Perceptual data across modalities (facial expression, voice, posture) is combined with structured knowledge about user identity, diagnosis, and interaction preferences. The world model is updated dynamically during interaction—for example, when a user changes location or emotional state—enabling the system to maintain temporally coherent representations of the user and the context.

The reasoning engine interfaces with this model to determine appropriate behaviors. It applies a set of symbolic rules informed by emotional state, user history, and environmental configuration. When perception is inconclusive, fallback reasoning leverages past observations and scene context to infer likely emotions and select suitable strategies. For instance, if a user's face is unreadable but they are alone in a clinical room with a history of anxiety, the robot may infer a fearful state and respond empathetically.

This capacity for reasoning under uncertainty, adapting to dynamic environments, and personalizing behavior demonstrates the system's core social intelligence. By tightly integrating perception, memory, symbolic reasoning, and multimodal actuation, TiagoCare enables human–robot interactions that are responsive, trust-building, and contextually grounded.

4 Implementation

This section details the implementation of the system's main functional modules, which collectively enable the TIAGo robot to perceive, interpret, and respond to the emotional and contextual states of users in real time. Each module is implemented as an independent ROS node and contributes to a modular and extensible architecture. We describe the technical design, data structures, and processing pipelines of the individual components — ranging from emotion recognition and context detection to behavior selection and simulation. Emphasis is placed on the integration of perception, reasoning, and actuation, with references to key libraries, tools, and machine learning models used throughout the system.

4.1 Tools and Libraries

The implementation integrates multiple open-source libraries and frameworks to support real-time perception, context inference, and behavior execution within a ROS-based architecture. Here, we provide a compact reference of the main libraries and their functional roles across the system. Specific usage details are explained in the corresponding module sections that follow.

Table 1: Summary of Major Libraries and Their Functional Roles

Purpose	**Libraries / Tools**
ROS integration and messaging	`rospy`, `std_msgs`, `geometry_msgs`, `sensor_msgs`, `cv_bridge`, `play_motion_msgs`, `rospkg`
Facial emotion recognition	`OpenCV`, `TensorFlow/Keras`
Audio emotion recognition	`PyAudio`, `torchaudio` `transformers`, `huggingface_hub`, `librosa`
Pose emotion recognition	`MediaPipe`, `scikit-learn`, `joblib`
Multimodal fusion	`collections`, `json`, `os`
Context detection	`ultralytics`, `face_recognition`, `PyKEEN`, `torch`, `pandas`, `numpy`, `datetime`, `pickle`
Behavior selection and personalization	`json`, `os`, `datetime`, `subprocess`, `pyttsx3`
Simulation orchestration	`subprocess`, `argparse`, `json`, `os`, `time`, `sys`, `signal`

4.2 Emotion Recognition from Face

The *Emotion Recognition from Face* module enables the robot to estimate a user's emotional state based on their facial expressions. Like the other emotion

modalities, this module is implemented as a standalone ROS node and supports adaptive, affect-aware robot behavior in real time.
The system captures video input from the onboard webcam using OpenCV. A Haar cascade classifier is applied frame-by-frame to detect faces. Once a face is detected, its bounding box is extracted as a region of interest (ROI) and processed for emotion classification. The ROI is resized to 224×224 pixels, converted to RGB, normalized to the [0,1] range, and formatted into a 4D tensor for model input.
The emotion classifier is a MobileNetV2-based neural network [21], fine-tuned on the FER2013 dataset [24], which outputs a softmax probability distribution over seven classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The model is loaded at runtime using TensorFlow/Keras. If the pretrained model is not found, the node automatically initializes a MobileNetV2 backbone with ImageNet weights [7] and appends a new classification head for robustness—though retraining is recommended for reliable performance.
To improve alignment during interaction, the module also includes a proportional head tracking controller. The center of the detected face is compared with the screen center, and a Twist message is published to adjust the robot's head position, keeping the user in frame.
The predicted emotion label (e.g., `"happy"`, `"sad"`) is published to a dedicated ROS topic as a std_msgs/String message. Emotion predictions are also logged in a local JSON file (emotions_face.json) to serve as a lightweight memory of the user's affective history.
This module was developed in Python and integrates both custom components, as ROS node, real-time face tracking, image preprocessing, fallback model logic, emotion logging, and snapshot saving; and external components, as MobileNetV2 architecture (via Keras), Haar cascade face detector (OpenCV), pretrained weights (FER2013 or ImageNet).

(a) Angry (b) Disgust (c) Fear (d) Happy (e) Sad (f) Surprise (f) Neutral

Figure 2: Examples of the seven emotion classes in the FER2013 dataset.

4.3 Emotion Recognition from Audio

The *Emotion Recognition from Audio* module is responsible for capturing live audio from the user's environment and analyzing it to infer the speaker's emotional state. This functionality is implemented as a standalone ROS node and contributes to the overall affect-aware behavior of the robot.
To acquire real-time audio data, the module uses the PyAudio interface

to access the microphone stream, initially sampling at 44.1kHz across two channels. Audio is processed in two-second segments, which are normalized, converted to mono, and downsampled to 16kHz using a resampling function. This preprocessed waveform is stored as a tensor, forming the input for emotion classification. These tensors serve as short-term auditory memory and enable low-latency inference.

The emotional inference is carried out using a deep learning model based on the Wav2Vec2 architecture [4], specifically `r-f/wav2vec-english-speech-emotion-recognition`, which is hosted on the Hugging Face Model Hub and was trained on SAVEE, RAVDESS and TESS datasets [11, 15, 18]. . The node uses the `transformers` library to load both the model and feature extractor via the `AutoModelForAudioClassification` and `AutoFeatureExtractor` APIs. During inference, audio features are extracted and passed to the neural model, which outputs logits corresponding to predefined emotion classes. The final predicted label is determined by selecting the class with the highest score, followed by optional remapping to a simplified emotion taxonomy. For instance, labels such as `disgust` and `surprise` are mapped to `sad` and `happy`, respectively, to align with the system-wide emotion schema.

Once an emotion is inferred, it is stored in a persistent JSON log file (`emotions_audio.json`) to support affect history and offline analysis. Simultaneously, the emotion label is published as a simple ROS `String` message to the `/tiago/audio_emotion` topic, enabling other nodes in the architecture to react in real time.

From an implementation standpoint, the audio acquisition loop, preprocessing pipeline, model inference, ROS integration, and logging have all been developed specifically for this project. The classification model, however, is reused from the Hugging Face ecosystem, which provides access to high-performance pretrained models and ensures reproducibility.

4.4 Emotion Recognition from Body and Hand Pose

The *Emotion Recognition from Body and Hand Pose* module enables the robot to interpret emotional cues from non-verbal behavior, focusing on full-body posture and hand gestures. It serves as a visual modality that complements the audio and facial emotion modules, enhancing multimodal emotion detection and robustness. Like the other modalities, this component is implemented as an independent ROS node.

The module acquires live video input from a webcam or video file and processes each frame using MediaPipe Holistic[16], a library for full-body landmark detection. Specifically, it extracts 3D landmarks for the body (33 points) and both hands (21 points each), yielding a feature vector of 225 coordinates (75 points $\times$ 3 dimensions), flattened into a 1D NumPy array. This vector is used as the core data structure for classification and serves as a short-term visual memory. To classify emotions from posture and gesture, the system employs a K-Nearest Neighbors (KNN) classifier[10] trained offline on the BEAST

(a) Angry (b) Happy (c) Sad (d) Fear

Figure 3: Examples of the four emotions in the BEAST dataset.

dataset[9]. This dataset provides labeled frames for basic emotions (happy, sad, angry, fear) with corresponding annotated body and hand postures. The extracted feature vector is fed into the KNN model for classification, and the numeric prediction is decoded back into a human-readable label using a LabelEncoder.

The pipeline runs continuously at 30Hz, extracting landmarks, predicting the emotion, and publishing the result as a std_msgs/String message on the ROS topic /tiago/pose_emotion. Simultaneously, the result is appended to a local JSON file (emotions_pose.json) to support temporal affect tracking and offline evaluation. If landmarks are not visible (e.g., occlusion or poor camera angle), the system defaults to a neutral prediction to maintain graceful degradation.

For debugging and user feedback, the module also draws detected landmarks on the video feed using MediaPipe drawing utilities. If the –quiet flag is not set, this visualization is rendered in a real-time OpenCV window titled "Pose and Hands."

4.5 Multimodal Fusion

Following the development of three distinct emotion recognition modules — based on facial expressions, vocal signals, and body posture — we introduce the *Multimodal Fusion* component. Its primary purpose is to integrate the predictions from these perceptual channels into a single emotional interpretation. This process determines the predominant emotional state of the user and consequently guides the robot TIAGo's socially aware behavior. By aggregating information from diverse sources, the system gains a more reliable understanding of the interlocutor's affective state, enabling it to select appropriate responses aimed at improving the human's emotional condition.

From a technical standpoint, the fusion mechanism operates by reading three JSON-formatted files (`emotions_face.json`, `emotions_audio.json`,

`emotions_pose.json`) — each corresponding to the output of a specific emotion recognition module. Each file contains a list of dictionaries, where each dictionary includes an `"emotion"` key associated with the predicted emotion for a specific frame. To ensure consistency across modalities and enable reliable comparison, a normalization step is applied. This step maps heterogeneous emotion labels to a unified taxonomy consisting of five classes: {`"angry", "happy", "sad", "fear", "neutral"`}.
Once normalization is complete, all valid emotion predictions are combined into a single list, which is processed using a frequency-based majority voting algorithm. This approach identifies the most frequently occurring label as the fused emotional state. The implementation includes a refinement, aimed at suppressing the `"neutral"` emotion: namely, if the label `"neutral"` is the most frequent but at least one other non-neutral emotion is present, the system discards `"neutral"` in favor of the next most frequent emotion. This strategy helps ensure that emotionally salient signals are not overshadowed by ambiguous or low-intensity cues.
The fusion logic is encapsulated in a dedicated ROS node named `emotion_fusion_node`. This node is designed to execute once per invocation: it loads the input files, performs normalization and fusion, and then publishes the resulting emotion as a `std_msgs/String` message on the topic `/tiago/fused_emotion`. Additionally, the output is saved to a file named `fused_emotion.json` to support asynchronous access or logging. Moreover, basic exception handling is included to manage the absence or corruption of input data. If one or more modality-specific files are missing or malformed, the fusion process continues using the remaining valid inputs, while issuing appropriate ROS warnings to notify the user.

4.6 Context Detection

The *Context Detection* module is responsible for enabling the robot to perceive and interpret its environment in a socially meaningful way. It integrates multimodal perceptual input with symbolic reasoning to infer not only the physical context in which the robot operates, but also the emotional and identity-related aspects of its human interlocutors. This module is implemented as a ROS node, `SceneTrackerNode`, which coordinates the real-time acquisition, interpretation, and encoding of contextual data.
Visual information is acquired through the robot's onboard camera and processed using a YOLO-based object detection model [20]. This allows the system to detect semantically relevant objects such as people, chairs, tablets, and laptops, which serve as proxies for activity and environmental context. In parallel, the system performs face recognition using a lightweight encoder from the `face_recognition` library. Detected face regions are converted into 128-dimensional embeddings and compared with stored user data. If a match is found, the associated identity is retrieved; if no match exists, a new identity is generated and added to a persistent user database. A short-term memory buffer ensures that identification remains consistent across frames, reducing flickering

due to transient misdetections.
After perception, the system constructs a structured representation of the scene in the form of semantic triples. These triples, expressed as `(subject, predicate, object)`, capture key environmental, emotional, and contextual relationships such as a person's location, current emotion, or their interaction with the robot. The resulting collection of triples constitutes the scene graph, which is persistently stored in `scene_graph_log.json`. Each entry is timestamped and associated with a specific user, allowing the robot to maintain a time-aware history of symbolic observations. For example, a log entry might contain `["person_1", "feels", "anxious"]` and `["person_1", "is_in", "clinic_room"]`. In parallel, a simplified version of the scene is recorded in `scene_log.json`, capturing flat key-value entries such as identity, location, and emotion for quick temporal reference.
The contextual information gathered from the scene is organized into a knowledge graph composed of semantic triples, which is then embedded into a continuous vector space using the TransE algorithm via the PyKEEN library. This embedding allows the system to perform symbolic reasoning and infer likely emotional states even when explicit emotion labels are unavailable. For instance, consider a case where the robot observes that a user is located in a `clinic_room`, is `sitting` on a `chair`, and is `interacting_with` the robot, but no direct emotional input is detected. To infer the user's emotional state, the system evaluates a set of candidate triples such as `("person_1", "feels", "anxious")` and `("person_1", "feels", "calm")` by computing plausibility scores within the embedding space. Specifically, it calculates the vector distance between the sum of the subject and relation embeddings and each candidate emotion embedding. If

$$\|\text{person\_1} + \text{feels} - \text{anxious}\| = 0.82 \quad \text{and} \quad \|\text{person\_1} + \text{feels} - \text{calm}\| = 1.34,$$

the system ranks `"anxious"` as more plausible than `"calm"` and selects it as the most likely affective state. Contextual cues—such as the location being a clinical setting—can also modulate these predictions, increasing the likelihood of emotions like anxiety in such environments. This form of context-aware emotional reasoning complements direct sensor-based emotion detection and enables more robust affective inference. Additionally, the system supports incremental learning: new observations can be added to the graph, and the model can be fine-tuned to reflect updated contextual patterns, allowing it to adapt to new users and dynamic environments over time.
Moreover, in cases where the user is new or has not been observed under the current conditions, the module supports incremental learning by fine-tuning the embedding model with additional triples. This allows the system to adapt to novel entities and maintain up-to-date representations without retraining from scratch. Finally, the predicted emotional state is published as a ROS message, which can be consumed by the robot's behavior planning system to select socially appropriate actions.

4.7 Social Reasoning

The *Social Reasoning* module is responsible for selecting the robot's behavior during interaction. It integrates input from multiple sources — including the user's emotional state, contextual information, and personal profile data — to determine appropriate and meaningful responses. In doing so, it enables the TIAGo robot to adapt its actions to the specific needs and situation of each user. The underlying process is managed by a ROS node that receives real-time data such as fused emotion predictions and semantic context triples. It consults any available user profile and issues symbolic commands to the components responsible for speech, gesture, and physical positioning. The system supports real-time interactions and it is capable of analyzing emotion trends over time. To express socially meaningful behaviors, the system utilizes three expressive channels — gestures, proximity, and verbal communication — each managed by an independent ROS node.

Gestures

A dedicated *gesture handler* was developed for TIAGo. This module receives symbolic gesture commands (e.g., `'wave'`, `'open_arms'`) via a ROS topic and translates them into predefined motions using the `play_motion` action server. Upon receiving a valid command, it maps the symbol to a motion name and constructs a `PlayMotionActionGoal`, which is then sent to the motion controller. This setup abstracts low-level motion control, allowing high-level components to trigger gestures in a modular and human-readable manner.

Table 2: Supported Gestures: Motion Name, Expected Output, and Intended Meaning

Motion Name	**Expected Output**	**Intended Meaning**
`wave`	Raises right arm and waves	Greeting or saying hello
`offer`	Opens arms forward	Invitation, empathy, support
`point`	Points with right hand	Directing attention
`thumb_up_hand`	Extends arm with thumbs-up	Encouragement, approval
`shake_hands`	Reaches out to shake hands	Greeting or farewell
`open`	Arms in open pose	Neutral readiness
`idle`	Returns to neutral posture	Rest or idle state
`nod`	Nods head gently	Agreement or acknowledgment

Proximity

The *proximity handler* allows TIAGo to adjust its physical distance from the user based on symbolic proximity levels—`close`, `medium`, and `distant`. The

module subscribes to the ROS topic `/tiago/proximity`, interprets the incoming command, and maps it to a linear velocity using a predefined dictionary. A corresponding `geometry_msgs/Twist` message is then published to `/mobile_base_controller/cmd_vel`, causing the robot to either move forward, remain stationary, or back away.

Table 3: Symbolic Proximity Levels, Motion Behavior, and Intended Meaning

Proximity	**Velocity (m/s)**	**Behavior**	**Intended Meaning**
close	+2.0	Robot moves forward to reduce distance	Engagement, empathy, or support
medium	0.0	Robot remains stationary	Neutral or respectful interaction distance
distant	-2.0	Robot moves backward to increase distance	Giving space or responding to discomfort

Voice

The *voice module* enables TIAGo to produce verbal responses by converting symbolic text messages into speech. It operates by subscribing to the `/tiago/tts` topic, where it receives input as plain text. Upon receiving a message, it uses the lightweight `espeak` text-to-speech engine, that is invoked through a system call to vocalize the content.

4.7.1 Behavior Selection Pipeline

All these modalities are coordinated through a centralized decision-making process. This behavior selection pipeline follows a three-stage process: rule-based inference, personalized adjustment, and behavior execution. This structure enables the robot to interpret user states and deliver contextually appropriate, socially expressive responses in real time.
The first stage is handled by the `infer_robot_action` function, which receives symbolic input in the form of semantic triples. These triples are systematically parsed to extract each user's emotional state, physical location, and any environmental or contextual tags. The function then evaluates this state against a set of deterministic rules that map the combined affective and contextual profile to a high-level robot behavior. The decision to condition certain emotional responses, specifically `"anxious"` and `"sad"`, on location context reflects the system's sensitivity to situational factors. For instance, anxiety in a waiting room may indicate anticipatory stress, prompting a warm, empathetic approach (`empathy_mode`), whereas the same emotion in a clinic room might be better addressed with a focused, low-complexity interaction (`simplified_dialog`). This context-dependent branching enables the robot to fine-tune its behavior to both emotional and environmental cues without requiring a full probabilistic model.

Other emotion-to-behavior mappings are direct and context-independent. A `"tired"` user triggers `low_energy_support`, while `"fearful"` users prompt `reassurance_protocol`. `"Anger"` leads to `conflict_diffusion`, and emotions such as `"excitement"` or `"curiosity"` result in `engage_with_enthusiasm` and `provide_information`, respectively.

In the second stage, the inferred behavior is refined through the `personalized_response` function, which enables user-specific adaptation based on stored interaction preferences. This function evaluates whether to retain the behavior inferred from contextual reasoning or to override it with a predefined, user-specific response stored in the patient profile database. In particular, the decision logic follows three main cases. First, if the user is new and has no meaningful stored preference (i.e., the profile contains only the system default, such as `monitor_only`), the system uses the inferred behavior without modification. Second, if the current diagnosis differs from the one stored in the user profile, the system assumes the profile may be outdated or inaccurate and again defers to the inferred behavior. Third, if the profile contains a diagnosis that matches the current one (or is unspecified), and includes a valid preferred response, this preferred behavior overrides the inferred action.

Finally, the selected behavior is passed to the `execute_robot_action` function, which translates the symbolic behavior class into multimodal robot actions. As shown in Table 4, each behavior corresponds to a specific combination of *gesture*, *speech*, and *proximity*. These actions are dispatched to dedicated ROS nodes that control voice synthesis, motion execution, and spatial positioning, allowing the robot to deliver coherent and responsive behaviors in real time.

Table 4: Behavior Classes and Associated Multimodal Outputs

Behavior Class	**Associated Actions**
`empathy_mode`	Speech: "You seem anxious. I am here to help you." Gesture: open_arms Proximity: close
`positive_reinforcement`	Speech: "I'm happy to see you in a good mood!" Gesture: thumb_up_hand Proximity: medium
`simplified_dialog`	Speech: "Let's keep it simple. How can I assist you today?" Gesture: wave Proximity: medium
`low_energy_support`	Speech: "Take your time." Gesture: wave Proximity: distant
`reassurance_protocol`	Speech: "Everything is okay. You're safe and I'm here with you." Gesture: offer Proximity: close
`conflict_diffusion`	Speech: "I notice some frustration. Let's work together." Gesture: open Proximity: distant
`engage_with_enthusiasm`	Speech: "You seem excited! Tell me more!" Gesture: thumb_up_hand Proximity: medium
`provide_information`	Speech: "I can help with that! What would you like to know?" Gesture: point Proximity: medium
`monitor_only`	Speech: (No action executed; monitoring only) Gesture: (None) Proximity: (None)
`unknown`	Speech: "I'm not sure how to respond, but I'm here." Gesture: (None) Proximity: medium

4.8 Simulation

The *Simulation* module manages the overall flow of the robot's interaction process. Implemented as the `TiagoSimulationNode` within the ROS framework, it integrates multiple subsystems — perception, context analysis, user modeling, and behavior execution — into a unified control loop. The module emulates realistic interaction sessions in which the TIAGo robot interprets multimodal signals and adapts its behavior accordingly. Upon initialization, the simulation node establishes communication with all relevant publishers and subscribers, ensuring that speech, gestures, and proximity controllers are ready for use.

The interaction begins with an initial greeting, designed to initiate engagement in a socially appropriate manner. Then, the emotion recognition process begins, structured into three sequential phases — each focused on a different modality: facial expressions, vocal cues, and body posture. These modules are launched as separate subprocesses and operate for a predefined time window, during which the robot provides guidance to extract meaningful user responses.

Following emotion detection, the simulation triggers the *Multimodal Fusion* component, which aggregates the individual outputs, performs emotion label normalization, and applies a majority-voting scheme to compute a final, dominant emotional state. The result is stored in a shared file for downstream access.

In parallel, environmental perception is carried out through the *Context Detection* module, which analyzes the robot's surroundings and encodes relevant spatial and semantic relationships as symbolic triples. These contextual representations serve as input to the reasoning process.

User recognition is carried out using either facial tracking or recent interaction data. If the user is not previously registered, the system creates a new profile and prompts for diagnostic input, which is entered manually via the `/tiago/new_user_input` topic and stored persistently. If the user has already interacted with the robot, their profile is retrieved automatically, enabling continuity across sessions.

Based on the fused emotional state, contextual information, and user profile, the reasoning engine (`infer_robot_action`) selects an appropriate high-level behavioral strategy. This decision is further refined through the `personalized_response` function, which adapts the response to the user's clinical background and interaction history. The resulting behavior is enacted through coordinated verbal and non-verbal actions, including speech, expressive gestures, and proximity modulation.

Throughout the interaction, the simulation node subscribes to two ROS topics — `/tiago/predicted_emotion` and `/tiago/new_user_input` — enabling dynamic updates and real-time adaptation. Once the session concludes, the robot delivers a short farewell, resets to a neutral state, and the simulation node terminates gracefully.

4.8.1 Illustrative Scenario

To illustrate the end-to-end operation of the *Simulation* module, consider the following example of a realistic interaction between the TIAGo robot and a user during a typical session:

1. **Greeting Phase:** Upon launch, the robot introduces itself by saying "Hello there! I'm TIAGo. I'm here to help and support you today," while performing a waving gesture and positioning itself at a medium interpersonal distance.

2. **Emotion Recognition:** The robot guides the user through three short affective assessment stages:

   - For facial analysis, the robot prompts: "Let me take a look at your face. Just look at me for a few seconds," accompanied by a pointing gesture.
   - For vocal emotion detection, the robot asks: "Now, please say something aloud. How are you feeling today?" while opening its arms.
   - For body posture analysis, the robot encourages playfulness: "Try putting your hands out in front of you, like a superhero!" with a thumbs-up gesture.

   During each phase, the respective emotion recognition node is activated for approximately 60 seconds.

3. **Multimodal Fusion:** Once all emotional cues have been collected, the robot announces: "Combining everything I sensed..." and executes the fusion module. The final emotion, for example `"sad"`, is extracted from the shared `fused_emotion.json` file.

4. **Scene Understanding:** The robot scans its environment, identifying semantic elements such as a person sitting on a chair in a living room. This information is encoded into symbolic triples and used to infer contextual relevance.

5. **User Identification and Profiling:** If the user is new (i.e., not previously seen), the robot asks: "Do you have a diagnosis or condition you want to share with me?" The user may respond via the ROS interface with input such as `"diagnosis: depression"`, which is saved under a persistent profile.

6. **Social Reasoning and Execution:** Based on the detected emotion `"sad"`, the environment (e.g., "person is sitting in living_room"), and the user's diagnosis, the reasoning engine might infer the action `"offer_encouragement"`. The robot then performs a behavior such as: "I'm here for you. Remember, you're not alone," coupled with a supportive gesture and reduced interpersonal distance.

7. **Session Closure:** The interaction concludes with the robot saying: "Thanks a lot! I'm always here if you need me. Take care!" while performing a handshake gesture and resetting to a neutral posture.

5 Results

The evaluation of TiagoCare confirmed its ability to recognize emotional cues from face, voice, and posture, and to respond in ways that are contextually appropriate and personalized. Despite occasional noise or misclassification, the multimodal fusion process remained generally stable and reliable. Personalized behaviors—driven by user profiles and emotional history—were activated in the majority of cases, and fallback reasoning mechanisms effectively handled inconsistencies, ensuring smooth interaction. The system also achieved strong results in context detection and user recognition, enabling consistent adaptation to the environment. Overall, these results highlight TiagoCare's potential as an emotionally intelligent assistive robot capable of integrating perception, reasoning, and social responsiveness.

Video-Based Simulation Overview

As described in Section 4.8, once the robot determines the dominant emotional state through multimodal fusion, it proceeds to context detection, user recognition, and behavior selection. The accompanying video demonstrates this pipeline by showcasing how TiagoCare tailors its response according to each user's emotional state and clinical profile.

Two contrasting examples are highlighted. In the first case, although `person_24` has a diagnosis of depression and a preferred response of `simplified_dialog`, the detected emotion ("happy") does not align with this diagnosis. Rather than rigidly following the profile, the system prioritizes the user's current mood and selects a more appropriate strategy: `positive_reinforcement`. The robot delivers a personalized message — "person_24, I'm happy to see you in a good mood" — accompanied by a thumbs-up gesture and stable medium proximity. The interaction concludes with a friendly farewell: "I'm always here if you need me. Take care!", reinforced by a handshake gesture.

In the second example, another user also classified as "happy" receives a distinct response based on their emotional history and clinical profile. In this case, `person_26` has a post-depression diagnosis and a preferred response of `engage_with_enthusiasm`. Since the detected mood aligns with the profile, the system applies the predefined strategy. The robot reacts enthusiastically: "person_26, you seem really excited! Tell me more!", highlighting its ability to reinforce positive emotional states in accordance with user preferences.

Finally, the video illustrates how the robot handles new users by initiating a data collection step: "Do you have a diagnosis or condition you want to share with me?". The user can respond via the terminal interface, and the system stores both affective data and contextual state for future interactions.

This simulation highlights how TiagoCare's reasoning and behavior modules work together to create a fluid and adaptive interaction loop — grounded in emotion recognition, symbolic reasoning, and user-specific personalization. The system not only adapts to the present emotional state but also knows when to prioritize real-time affect over stored diagnostic assumptions.

Goal Achievement

As outlined in Section 1.2, the objectives of this project were threefold: (1) perceive human emotional states in real time using multimodal input, (2) reason symbolically over context and personal history to drive behavior, and (3) personalize robot interaction to support emotional well-being. Our results confirm that:

- Multimodal Emotion Recognition was successfully implemented through face, audio, and pose classifiers. Despite moderate accuracy, the fused emotion pipeline ensured reliable emotion interpretation in most interactions.
- Symbolic Reasoning via scene and emotion graphs enabled context-aware behavior selection. Personalized strategies were activated based on semantic triples like `(person_x, context_state, medical_anxiety)`.
- Personalization and Adaptivity were supported through dynamic user profiles. Stored data such as diagnosis and past responses were used to guide fallback behavior even when perception failed.

5.1 HRI

In addition to the scenarios previously discussed, we now analyze four of the remaining test cases in more depth, focusing on two key aspects: (1) whether the robot responds as expected based on the user's clinical profile, and (2) how effective the robot's reasoning abilities are in adapting to the interaction context.

5.1.1 Test Users Overview

To assess the effectiveness of the reasoning and behavior modules implemented in TiagoCare, we conducted a controlled user study using synthetic participants modeled after real clinical profiles. Seven test users were simulated based on the `patients_db.json` profiles. Each participant was assigned a diagnosis, response pattern, and preferred robot behavior (Table 5). The aim was to determine whether personalized, diagnosis-driven behavior led to improvements in user experience—particularly comfort, emotional alignment, and perceived empathy—compared to default robot responses.
Note: During the simulation, additional synthetic identities (e.g., `person_24`) may appear. These users were used exclusively for internal testing and debugging purposes and are not part of the main evaluation set. Only the seven profiles listed in Table 5 were included in the structured user study and qualitative assessment reported in this section. The following table summarizes the full set of test users showing their diagnosis, expected emotional behavior, and preferred robot response strategy.

Table 5: Synthetic User Profiles Used for HRI Evaluation

User	Diagnosis	Response Pattern	Preferred Response
person_1	Social Anxiety	Avoids eye contact	Simplified Dialogue
person_2	Depression	Benefits from positive reinforcement	Positive Reinforcement
person_3	None	Standard	Monitor Only
person_4	Depression	Prefers to be reassured	Reassurance Protocol
person_5	General Anxiety	Responds well to calm gestures	Empathy Mode
person_6	Post-Op Stress	Needs simplified explanations	Simplified Dialogue
person_7	Social Anxiety	Prefers minimal interaction	Monitor Only

5.1.2 Reasoning Behavior and Adaptation Analysis

To better understand the robot's ability to reason and adapt, we report key examples of how the reasoning pipeline reacted in live simulations:

Neutral Misclassification (person_5): Although person_5 (general anxiety) showed fearful posture and voice cues, the fused emotion occasionally resulted in a 'neutral" label. However, the reasoning module used the patient's diagnosis and historical emotion trend (fearful episodes in recent logs) to override the neutral prediction and selected `empathy_mode`, performing a calm gesture and using supportive speech.

Diagnosis Mismatch Detected (person_1): In this case, the stored profile labeled `person_1` as having social anxiety, but the current input indicated a different diagnosis. The reasoning module, via the `personalized_response` function, detected the mismatch and opted for the inferred response instead of the potentially incorrect profile, preserving safety and adaptability.

No Interaction Preferred (person_7): Despite person_7 occasionally being misclassified as mildly anxious, the robot respected the 'monitor_only" policy tied to their social anxiety diagnosis. No gesture or speech was performed, and the robot maintained distant positioning, showing that the system prioritizes user-defined preferences over uncertain emotional predictions.

Consistent Execution (person_6): With a clear 'angry" fused emotion and a diagnosis of post-operative stress, the robot correctly selected the `simplified_dialog` strategy: short instructions, medium distance, and minimal gesturing. This matched both the inferred and preferred responses, validating the rule-based logic.

These case studies show that the robot's behavior is not only guided by emotion detection but also by a structured reasoning system that factors in profile data, diagnosis consistency, emotional history, and fallback mechanisms. This improves robustness in real-world uncertain conditions.

5.1.3 User Study

To evaluate the quality of interaction enabled by our HRI system, we designed a small-scale, simulation-based study centered on the robot's behavior with four specific users (`person_4` to `person_7`).

This part of the evaluation aims to understand whether the system's reasoning abilities—grounded in symbolic profiles and knowledge graph triples—can generate context-aware and emotionally aligned behavior, even when emotion detection is noisy or inconclusive. In particular, we wanted to observe whether personalization improved user comfort and empathy, and whether fallback strategies triggered correctly when the input was inconsistent. The study was structured around two interaction modes. In the control condition, the robot executed a standard, non-personalized script regardless of the user profile. In the experimental condition, it adapted its verbal and non-verbal behavior using diagnosis, preference, and emotion inputs. For each participant, we simulated both conditions, capturing fused emotions before and after interaction, and collecting synthetic feedback through Likert-scale responses and qualitative comments.

Our working hypothesis was that tailored responses—such as simplified dialogue for post-operative stress or reassurance for depression—would lead to better perceived interaction quality compared to generic interaction. At the same time, we remained cautious about the role of the emotion recognition module, which in our current implementation is still prone to misclassification, especially in complex or ambiguous cases like 'neutral".

Table 6 reports the outcomes of the interactions in personalized mode, focusing on the final detected emotion, the robot's selected strategy, and the simulated user feedback on comfort and perceived empathy.

Table 6: Feedbacks from Test Users (Personalized Condition)

User	`person_4`	`person_5`	`person_6`	`person_7`
Diagnosis	Depression	General Anxiety	Post-Op Stress	Social Anxiety
Fused Emotion	Sad	Neutral (expected: fear)	Angry	Neutral (stable)
Strategy Used	Reassurance Protocol	Empathy Mode	Simplified Dialogue	Monitor Only
Comfort (1–5)	4	4	4	3
Empathy (1–5)	4	3	3	2
User Comment	The robot's words were comforting.	It helped, although it didn't seem to notice how anxious I was.	I appreciated the clear instructions, but it still felt a bit generic.	It did nothing, which was okay—but maybe too passive.

On average, the comfort score across these four interactions was 3.75 out of 5, while perceived empathy reached 3.0. While these results confirm that the robot was generally able to respect user preferences and avoid disruptive behavior, they also highlight important limitations—especially in cases where the emotion fusion module failed to detect anxiety or fear signals, defaulting to a misleading 'neutral" label. This occurred in particular for `person_5`, where the robot executed the correct behavior thanks to fallback reasoning based on the user's stored diagnosis, but without explicitly recognizing the emotional state.
Our results suggest that the reasoning component and symbolic profile management in TiagoCare provide a useful backbone for adaptation in HRI. Even when multimodal emotion fusion is not reliable, the robot can fall back on profile-based decision-making to avoid inappropriate reactions. That said, the quality of emotional alignment still depends significantly on the accuracy of the input signals. In the case of `person_7`, the robot respected the user's preference for minimal interaction, but the experience was perceived as slightly too passive—suggesting the need for more nuanced behavior even within the "monitor only" category.
Nonetheless, further refinement of emotion recognition and dialogue dynamics is needed to reach higher levels of fluency and trust in real-world assistive environments.

5.2 RBC

The following section provides a structured evaluation of the TiagoCare system, following the RBC methodology. We assess both the performance of individual modules (functional benchmark) and the system's ability to carry out context-aware, socially appropriate interactions (task benchmark).
We also describe the world modeling and symbolic reasoning strategy that supports adaptation to user needs, even in ambiguous scenarios. Together, these benchmarks offer a comprehensive view of the system's technical capabilities and interaction quality.

5.2.1 Functional Benchmark

Emotion Recognition From Face The Emotion Recognition From Face module has been benchmarked on the FER2013[24] dataset, using accuracy as the evaluation metric. The model has been trained on a training set composed of 80% of the dataset, and evaluated on a test set composed of the remaining 20%. The results of the evaluation are reported in the image 4, reaching a final accuracy of 0.62.

Emotion Recognition From Audio The Emotion Recognition from audio leverages a pretrained Wav2Vec2-based model from the Hugging Face Model Hub, enabling real-time inference through a streamlined classification pipeline. The testing results of that model are reported from the hugging face model

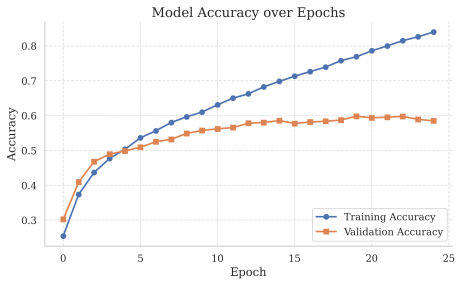

Figure 4: Benchmark Scores Across Trials For Face Emotion recognition

hub, achieving an accuracy of **0.97463** on the evaluation dataset composed of unforeseen data from TESS, RAVDESS and SAVEE datasets.

Emotion Recognition From Pose For the Emotion Recognition From Pose module we personally trained a KNN [10] on the BEAST[9] dataset. The evaluation of this model has been performed on a test set of unforeseen data coming from that dataset and using accuracy as an evaluation metrics. The evaluation has been performed on 7 different trials, in Table 7 are reported the results of the trained models.

Table 7: Benchmark Scores Across Trials For Pose Emotion recognition

Trial 1	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Trial 6**	**Trial 7**	**AVG**
0.696	0.677	0.701	0.682	0.724	0.702	0.697	0.697

Multimodal Emotion Fusion To fuse predictions from audio, pose, and face modules, we adopt a simple majority voting approach, selecting the most frequently detected emotion as the final fused label. In case of ties or ambiguity, the system falls back on context-based reasoning using user profiles and historical logs.

Context Detection We use YOLOv8 [20] for object detection, recognizing key environmental elements such as *chair*, *laptop*, *table*, and *tablet*. Detected objects are converted into scene graph triples (e.g., `person_1 sits_on chair_3`). The YOLOv8 model achieved an overall precision of 0.87, demonstrating robust performance in detecting relevant elements in indoor assistive environments.
For context reasoning, a set of 200 synthetic triples was created to represent semantic knowledge. We trained a TransE model using PyKEEN with these triples. The model converged after 87 epochs, reaching a final training loss of 0.032. This low loss value suggests that the model effectively learned to represent relations between users, environments, and emotional states in a structured and predictive way.

Importantly, the knowledge graph is designed to support continual learning. At runtime, when new observations are logged (e.g., a new user joins, or new emotional states are detected), the model is incrementally fine-tuned. This enables the robot to adapt its understanding of social and emotional context as the environment and user behavior evolve.

5.2.2 Task Benchmark: Interaction Evaluation

We defined a task as successful when the robot was able to:

- Correctly identify the user (via stored profile)
- Recognize emotional cues from face, audio, and pose
- Infer the dominant emotion through multimodal fusion
- Generate coherent semantic triples describing the state
- Perform an appropriate, context-aware response in gesture, speech, and spatial positioning

In our simulations, most of these criteria were satisfied in the majority of test runs. Cases such as `person_5` (misclassification as "neutral") demonstrate the robustness of the fallback reasoning system, which was able to infer a proper response based on context and user history even when the fusion module failed.

5.2.3 World Modeling and Context Reasoning

Our system adopts a symbolic world modeling strategy based on scene and emotion graphs, designed to support context-aware and socially adaptive behavior. The world is represented through RDF-style triples—(subject, predicate, object)—generated dynamically at runtime and stored in a local knowledge base. These triples, as previously mentioned, fall into three main categories: scene, emotion, and contextual triples. This structured representation enables TIAGo to adapt its behavior based on both immediate sensory inputs and longer-term user history. Compared to prior HRI systems focused solely on perception or unimodal processing [3], our approach integrates symbolic reasoning directly into the interaction loop.
Inspired by recent work on affective computing [19] and [13], we demonstrate that TransE-based embeddings can generalize across similar users and environments, even when input signals are noisy or ambiguous. For example, given the input:

```
(person_1, is_in, clinic_room)
(person_1, feels, anxious)
```

the system can predict:

```
(person_1, context_state, medical_anxiety)
```

which in turn triggers a tailored response such as: *"You seem anxious. I am here to help you."* This integration of symbolic context with affective reasoning allows TIAGo to adapt its behavior not only based on current emotion, but also by reasoning over diagnostic profiles, recent emotion history, and spatial cues. By combining this with multimodal fusion and user memory, the system achieves a form of grounded empathy that moves beyond pre-defined scripts, offering flexible social interaction strategies.

5.2.4 Contribution

Our work both confirms existing findings and introduces new perspectives within the field of affective human–robot interaction. On one hand, our results validate what has been consistently observed in the literature: that multimodal emotion recognition—when combining face, voice, and body posture—can lead to more accurate and resilient affect detection compared to unimodal approaches. Benchmarks on FER2013 (face), TESS, RAVDESS, and SAVEE (audio), and BEAST (pose) datasets confirm the reliability of our perceptual modules, and our fusion strategy demonstrates its utility in handling ambiguous or partially missing signals. However, we also propose a novel integration of these perceptual capabilities with symbolic reasoning and world modeling. Unlike systems that rely solely on end-to-end neural networks or predefined interaction rules, TiagoCare leverages structured knowledge in the form of dynamically generated triples and TransE embeddings to reason about user context and emotional history. This allows the robot to go beyond reactive behavior and instead perform context-aware adaptation—taking into account not only what the user feels, but where they are, what diagnosis they carry, and how they have responded in the past. This hybrid architecture builds upon the ideas proposed in recent literature on context-aware emotion recognition by [19] and [13], but diverges in key aspects. Most notably, we demonstrate that lightweight symbolic reasoning, grounded in real-time data and incrementally updated, can serve as a powerful fallback strategy when perception fails. Our system is able to infer latent emotional states (such as 'medical anxiety') from environmental and diagnostic cues and adjust its verbal, gestural, and proxemic behavior accordingly. In this sense, TiagoCare does not simply confirm existing models—it extends them by embedding emotional perception within a continuously evolving symbolic context. By doing so, it highlights the potential of combining deep learning with knowledge graphs in assistive robotics, where personalization, interpretability, and robustness are essential.

6 Conclusion

The development of the TiagoCare framework has been a highly engaging experience. By combining multiple AI modules — from emotion recognition and symbolic reasoning to multimodal fusion and behavioral adaptation — we have gained hands-on insights into the complexities and potential of socially-aware assistive robotics. Through this project, we learned how to design and integrate perception pipelines for different modalities (face, voice, pose), structure world knowledge using scene graphs, and build rule-based reasoning mechanisms to personalize human–robot interaction. We also appreciated the importance of robust ROS-based communication and the challenges involved in simulating realistic, emotionally sensitive behavior. Despite these achievements, we acknowledge that several aspects of the system could be improved. First, the accuracy of individual perception modules, especially in ambiguous or noisy contexts, needs to be improved through better training data, fine-tuning, or use of more advanced models. Second, the system currently handles a limited set of emotions and predefined user profiles. Adapting it to more diverse users and real-world clinical environments would require more scalable and adaptive mechanisms, possibly incorporating learning-based dialogue, dynamic knowledge updates, and more expressive interaction strategies. Additionally, the reasoning module could be extended to support probabilistic or hybrid (neuro-symbolic) inference. Beyond technical improvements, this project also raises important non-technical considerations. Ethically, the deployment of emotionally aware robots in healthcare contexts must be guided by clear consent, transparency, and privacy principles. Psychologically, one must ensure that such systems do not create false expectations or emotional dependency, particularly among vulnerable populations. From a societal and economic perspective, it is crucial to reflect on the implications of partially replacing human caregivers with robotic assistants — and to ensure that such technology supports, rather than displaces, human care.

References

[1] Human-Robot Interaction 2025. Lecture slides.

[2] Robot Benchmarking and Competitions 2025. Lecture slides.

[3] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(01):73–79, 2021.

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 06 2020.

[5] Felipe Cid, Luis J Manso, and Pedro Núnez. A novel multimodal emotion recognition approach for affective human robot interaction. *Proceedings of fine*, pages 1–9, 2015.

[6] Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. Ari: The social assistive robot and companion. In *2020 29th IEEE International conference on robot and human interactive communication (RO-MAN)*, pages 745–751. IEEE, 2020.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[8] Sergio Garcia, Francisco Gomez-Donoso, and Miguel Cazorla. Enhancing human–robot interaction: Development of multimodal robotic assistant for user emotion recognition. *Applied Sciences*, 14(24):11914, 2024.

[9] Beatrice Gelder and Jan Van den Stock. The bodily expressive action stimulus test (beast). construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. *Frontiers in Psychology*, 2:181, 08 2011.

[10] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In Robert Meersman, Zahir Tari, and Douglas C. Schmidt, editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 986–996, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[11] Philip Jackson and Sana ul haq. Surrey audio-visual expressed emotion (savee) database, 04 2011.

[12] Ray Jarvis. Multimodal robot/human interaction in an assistive technology context. In *2009 Second International Conferences on Advances in Computer-Human Interactions*, pages 212–218. IEEE, 2009.

[13] Zihan Lin, Francisco Cruz, and Eduardo Benitez Sandoval. Self context-aware emotion perception on human-robot interaction. *arXiv preprint arXiv:2401.10946*, 2024.

[14] Zhen-Tao Liu, Fang-Fang Pan, Min Wu, Wei-Hua Cao, Lue-Feng Chen, Jian-Ping Xu, Ri Zhang, and Meng-Tian Zhou. A multimodal emotional communication based humans-robots interaction system. In *2016 35th Chinese Control Conference (CCC)*, pages 6363–6368. IEEE, 2016.

[15] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018.

[16] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019.

[17] Lorena Muscar, Lacrimioara Grama, and Corneliu Rusu. Sound classification by the tiago service robot for healthcare applications. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2021.

[18] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (tess). Borealis Data Repository, 2020.

[19] Katie Powell, Sarath Kodagoda, and Teresa Vidal-Calleja. Towards context aware emotion recognition in hri for social robots. In *Australasian Conference on Robotics and Automation, ACRA*, 2023.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[21] Mark Sandler, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. pages 4510–4520, 06 2018.

[22] Sudhir Shenoy, Yusheng Jiang, Tyler Lynch, Lauren Isabelle Manuel, and Afsaneh Doryab. A self learning system for emotion awareness and adaptation in humanoid robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 912–919. IEEE, 2022.

[23] Mohamed Ala Yahyaoui, Mouaad Oujabour, Leila Ben Letaifa, and Amine Bohi. Multi-face emotion detection for effective human-robot interaction. *arXiv preprint arXiv:2501.07213*, 2025.

[24] Lutfiah Zahara, Purnawarman Musa, Eri Prasetyo Wibowo, Irwan Karim, and Saiful Bahri Musa. The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi. In *2020 Fifth International Conference on Informatics and Computing (ICIC)*, pages 1–9, 2020.

[25] Hongbo Zhu, Chuang Yu, and Angelo Cangelosi. Explainable emotion recognition for trustworthy human–robot interaction. In *Proc. Workshop Context-Awareness Hum.-Robot Interact. Approaches Challenges ACM/IEEE HRI*, 2022.