Future Technology
Open Access Journal
ISSN 2832-0379

May 2026| Volume 05 | Issue 02 | Pages 200-210

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.2.19

Future Publishing LLC

Article

Emotionally intelligent social robot for dementia care: empathy-based conversational intervention model using multimodal sentiment analysis

Zhenyu Lei[1]*, Yiqiao Yin[2], Yingsheng Chen[1]

[1]School of Health Care and Nursing, Tianfu College of Swufe, China
[2]Financial Section II, Neijiang Municipal Bureau of Finance, China

ARTICLE INFO

Article history:
Received 05 November 2025
Received in revised form
09 January 2026
Accepted 05 February 2026

Keywords:
Multimodal sentiment analysis,
Socially assistive robots, Dementia care,
Empathetic dialogue generation,
Affective computing

*Corresponding author
Email address:
519602119@qq.com

DOI: 10.55670/fpll.futech.5.2.19

ABSTRACT

Dementia is a challenging health issue for healthcare systems across the globe. Communication disabilities and behavioral changes have been significantly impacting the well-being of patients with dementia. The study formulates an empathy-based conversational intervention approach for patients with dementia using multimodal sentiment analysis. The proposed system applies a cross-modal attention-based model to synthesize speech, facial expressions, and biological signals for effective emotion identification. The synthesis is further augmented with an innovative large language model-based conversational response generation module that can develop appropriate empathetic responses. Experiments conducted on public benchmark databases confirm that the trimodal fusion-based model outperforms state-of-the-art methods with an overall weighted average accuracy of approximately 87.3% for emotion identification. The proposed approach outperforms state-of-the-art methods such as MulT, MISA, and MAG-BERT. The scores on human evaluation of the generated empathetic dialogue reached 4.12 and 4.28 in terms of empathy and coherence, with improvements of 17.0% and 12.3% over baseline models. The meta-analytic synthesis of previous clinical evidence revealed significant beneficial effects of social robot interventions on depression, loneliness, and agitation of people with dementia. The comparison with commercial models such as PARO, Pepper, and NAO showed the superiority of the proposed approach over others in terms of multimodal emotion recognition and dialogue adaptability. These results show that socially interactive robots with high emotional intelligence, equipped with cutting-edge affective computing and natural language processing, have great potential for enhancing the quality of dementia care through personalized emotional support.

1. Introduction

The rise in the global incidence of dementia has been unprecedented, making it a focal point in public health in the twenty-first century. Statistics from the Global Burden of Disease Study 2019 revealed that 57 million people worldwide are living with dementia. Projections suggest that this number will rise to 153 million by 2050 [1]. The Lancet Commission on dementia prevention, intervention, and care emphasizes the need to develop new approaches to address the complex presentations of dementia that include cognitive decline, neuropsychiatric changes, and communication difficulties [2]. Apart from the health issues faced by patients, the financial toll of dementia on the healthcare systems is also quite considerable on a global level, with forecasts showing that the economic burden on the healthcare infrastructure of 152 countries worldwide will continue to intensify in the coming years [3]. The link between social factors and the rapid progression of dementia has long been a major concern for many researchers, with conclusive evidence pointing to a close relationship between social isolation, loneliness, and rapid dementia progression in said groups [4]. In a rather current systematic review, there has been a clear emphasis on loneliness having a remarkably high prevalence in people suffering from either mild cognitive impairment or dementia [5], while studies in the realm of cognitive aging continue to emphasize the mutual relationship between social isolation and individual cognition [6]. Challenges faced by people living with dementia can be seen reflected in the problems faced by carers, who suffer from intense psychological and physical distress. Carers have reported increased levels of loneliness

and isolation, starting a chain reaction of disturbances in the quality of caring [7]. Issues of carer burden encompass a range of problems, including emotional depletion, poor health, meaning the need for well-integrated support systems [8]. Online supportive interventions have proved their effectiveness in lessening levels of carer distress [9], while meta-analytic evidence has shown considerable reductions in carer distress from specific interventions for carers of people living with dementia [10]. Communication represents a basic challenge in dementia, where the enhancement of effective communications has emerged as a grand challenge requiring new solutions [11]. There have been positive impacts of communication approaches on outcomes related to the quality of life from systematic reviews [12], while approaches centered on languages, focusing on person-centered communications, have proved effective in more formalized caring environments [13]. These have very high human resource requirements, making scaling-up more complicated, especially in a situation where the demand for more nursing staff increases. Social robots can be a technological method of meeting this challenge. There has been considerable interest in the potential advantages of robots in enabling safe human-robot interaction for vulnerable individuals [14], with the concept of socially assistive robot technology having been identified in the context of elderly care by scoping reviews [15]. There has been evidence of meta-analysis implying the potential advantage of social robots in terms of the enablement of beneficial outcomes for users [16], having been supported by recent evidence signifying major developments in the reduction of depression and loneliness within long-term care facilities [17].

Theoretical bases of socially interactive robots that possess emotional intelligence are based on affective computing, which has developed extensively in recent years [18]. The implementation of technology for older people remains an essential point of consideration because meta-analysis has identified that perceived usefulness and perceived ease of use are strong antecedents of their intention to use technology [19]. The multimodal analysis of emotions has effectively tackled the constraints of single-modal methods of recognizing emotions. Thorough reviews of research papers prove significant improvement in joint processing of text, acoustics, vision [20], and deep learning surveys confirm the efficiency of attention methods in capturing multivariate relationships [21]. Finally, the development of empathetic conversation AI systems marks a key milestone, since studies have shown the ability of artificial systems to produce supportive responses suitable for a given context [22]. Systematic reviews of the empathetic pros and cons of a large language model have been positive, although pointing towards key limitations [23]. Despite these breakthroughs, social robots still demonstrate remarkable challenges in emotion recognition, particularly when engaging with people with dementia, who use distinctive communication patterns due to their cognitive impairment and emotional instability. State-of-the-art emotion recognition models are largely based on single-modal sentiment analysis, which does not effectively capture subtle emotional information expressed through multiple modalities simultaneously. The combination of multimodal emotion recognition and empathetic dialogue synthesis in social robots targeting people with dementia has remained significantly uncharted. This research fills these essential gaps by considering three related aims: to create an emotion analysis system that fuses speech, facial expression, and physiological data, and is adjusted for communication style to match individuals with dementia; to build an empathy-driven conversational treatment approach for producing emotionally supportive messages; and to evaluate the efficacy of the proposed system by using publicly available benchmark corpora and secondary meta-analytic analysis. This research is expected to provide theoretical contributions to understanding human and robot emotion interactions in cognitively impaired people, conceptual contributions by proposing a unique fusion approach of cross-modal attentions, and application-level contributions by providing a scalable technological approach for improving dementia care quality and alleviating caregiver workload worldwide.

2. Research methodology

2.1 Research design

In the proposed work, computational modeling and secondary data analysis approaches are combined in a mixed-method study to design and validate an empathy-based chat intervention system targeted towards people with dementia. The qualitative component involves thematic analysis of interaction patterns derived from dialogue transcripts and content analysis of system-generated responses, which complements the quantitative computational modeling by providing interpretive insights into the emotional dynamics of human-robot conversations. The proposed framework is represented using a five-layer modular design structure shown in Figure 1.

The first layer receives four different modalities of input: speech signal input, facial expression input, physiological signal input, and automatic transcription of speech (ASR text) input. The automatic speech recognition (ASR) transcript is derived from the speech signal through automatic transcription, acting as a supplementary textual representation with embedded semantic information, whereas the speech signal carries the relevant acoustic and prosodic information. These modalities are then processed by different feature extraction models, namely Wav2Vec2.0 for speech signal processing, ResNet50 for vision processing, 1DCNN-BiLSTM for physiological signal processing, and BERT for textual information processing. Wav2Vec2.0 has been chosen for speech signal processing due to its high accuracy in low-resource and noisy audio conditions, which are common in care facilities. ResNet50 has been chosen for vision processing over Vision Transformer models due to its efficiency and accuracy in processing small datasets, which is important in the absence of dementia-related facial expression datasets. BERT has been chosen for textual information processing due to its high contextual understanding capabilities, and 1DCNN-BiLSTM has been chosen for physiological signal processing due to its ability to learn both short- and long-term dependencies in time series physiological signals. In order to reduce the intrinsic asynchrony of the multimodal data, all the data are resampled to a common time resolution of 25 frames per second, with the speech segments being aligned with the corresponding facial expression frames through timestamp synchronization. The physiological signals are downsampled and aligned using linear interpolation to match the frame rate of the visual data.

The cross-modal fusion layer employs a self-attention mechanism in combination with the cross-modal fusion mechanism to effectively monitor intra-modal as well as inter-modal associations. The fusion mechanism employed in this layer is composed of a four-layer transformer encoder with eight attention heads and a hidden dimension of 256. In this layer, self-attention is employed to effectively capture intra-modal associations in each modality.

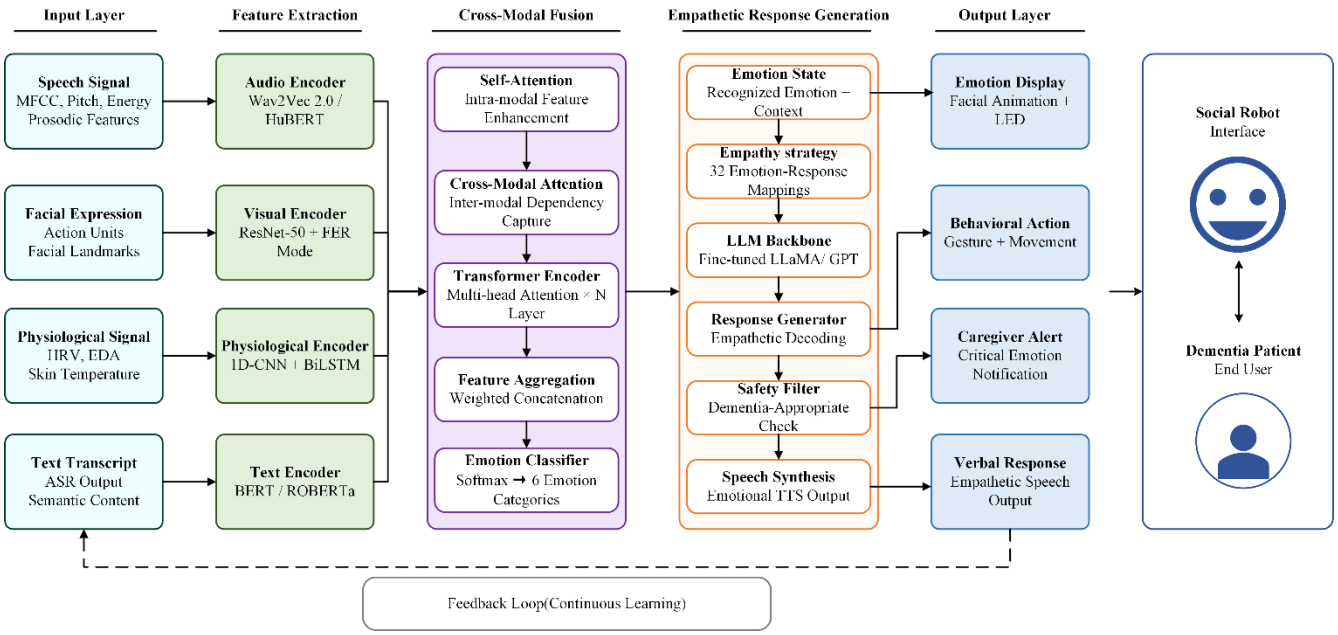

Figure 1. Architecture of the empathy-based multimodal emotion recognition and conversational intervention system for dementia care

Then, cross-attention is employed to capture inter-modal associations. The fusion weights are calculated dynamically based on the confidence scores of each modality. This architecture will allow a transformer classifier to recognize six different emotional states for each input type. The six emotional states will be anger, happiness, sadness, fear, surprise, and neutral. These emotional states have been selected in accordance with Ekman's basic emotion theory and have been found to be prevalent in dementia patients, where negative emotions such as sadness and anger are often caused by cognitive frustration, and fear is often caused by unfamiliar surroundings and confusion about daily activities. The response generation module that utilizes empathy makes use of decision logic based on the identified emotional categories, which are based on creative strategy mappings. This utilizes a highly tuned large language model to generate an appropriate response, which then goes through a safety filter. The final module produces responses in terms of verbal communication, facial expression, behavioral action, and notification messages based on the identification of the key emotional categories. Verbal responses are generated by an emotional text-to-speech system that modulates pitch, speech rate, and tone to match the target emotional characteristics. Facial expressions are generated by a parameterized avatar system that supports the six basic emotions. Behavioral actions include gestures and proximity adjustments appropriate to the emotional context. Caregiver notifications are triggered when the system detects critical emotional states such as persistent sadness, fear, or agitation exceeding predefined thresholds.

2.2 Multimodal sentiment analysis module

Multimodal sentiment analysis is the basis of the emotion recognition part of the structure, using speech, facial, and physiological information to identify the complex spectrum of emotions experienced by dementia sufferers. The selection of these three modalities is particularly suited for dementia populations. Speech analysis captures the prosodic changes and verbal hesitations commonly observed in cognitive impairment.

Facial expressions provide nonverbal cues that remain relatively preserved even when verbal abilities decline. Physiological signals offer involuntary indicators less susceptible to the communication difficulties experienced by dementia patients, as documented in studies examining emotional expression patterns in neurodegenerative conditions. Systematic reviews show that the hybrid models, using a combination of convolutional and recurrent neural networks, enable credible results in speech emotion recognition, valid for a wide range of acoustic conditions [24]. Graph neural networks have also been proven highly efficient in modeling complex temporal dependencies found in speech-related emotion signals, based on Mel-Frequency Cepstral Coefficients, Pitch Contours, and Prosody features [25]. However, applying the method to elderly people suffering from cognitive impairment is fraught with challenges. There have been a number of recent systematic reviews of key elements which influence generalization capability in elderly care facilities while using deep learning methods [26]. In light of changes occurring due to old age on face features, new, specifically Alzheimer's disease-related solutions have shown promising results, including dynamic processing pipelines involving face detection and localization, as well as disease-specific features [27]. Speech patterns in dementia populations present unique challenges including reduced speech rate, increased pause duration, and word-finding difficulties that differ markedly from typical adult speech corpora. Recognition accuracy has been reported to drop by 15-25% when models trained on standard datasets are applied to elderly speakers with cognitive impairment.

In comparison to facial and speech-related cue systems, physiological signs provide information regarding emotions that is more difficult to control. In terms of cardio-related emotion perception, there have been many reviews that are beneficial for discussing processes involved with respect to HRV and EDA signal feature identification [28] and helpful for creating multi-modal systems using data from wearable devices [29].

Integration of the heterogeneous modalities requires sophisticated fusion techniques for handling the asynchronous signals with varying levels of reliability. Multimodal transformers involving feature restoration on both levels have shown robust performance in situations where some of the modalities are absent or of poor quality [30], of practical relevance for a dementia-supporting environment. Fusion nets incorporating the attention mechanism have enhanced the ability for modeling interactions across modalities through adaptive weighting of the modalities based on their quality [31].

2.3 Empathy-based conversational model

The empathy-driven conversation method employs the use of identified emotional states and responds with appropriate and supportive messages targeting people living with dementia. The key element of the method is based on creating a relationship between identified emotional categories and their associated empathy methods, allowing the selection of appropriate responses based on the patient's primary emotional state. The empathy strategy mappings link each identified emotional state to appropriate response patterns. For sadness, the system employs emotional validation combined with gentle reassurance. For anger, the strategy involves acknowledging the emotion while redirecting attention to calming topics. Fear triggers responses that provide orientation and security cues. Happiness is met with shared positive affect and encouragement. Surprise prompts clarifying and grounding responses. Neutral states allow for general conversational engagement or activity suggestions based on patient preferences. Effective studies from recent research work have managed to validate that large language models work effectively in providing empathetic responses through techniques such as prompt engineering [32]. The response generation module in this study utilizes LLaMA-2-7B as the base model, fine-tuned using Low-Rank Adaptation on a curated dataset combining EmpatheticDialogues with dementia care conversation transcripts from published clinical studies. Domain adaptation was performed by incorporating vocabulary patterns and communication strategies documented in person-centered dementia care literature, including simplified sentence structures and repetition-tolerant dialogue flows.

The dialogue management aspect of the module employs continuous state tracking. This feature allows the system to be aware of its conversation state and to track the affect state of every conversation exchange. Another aspect of the module's dialogue management is its personalization feature. The personalization feature of the system is made possible with the creation of a patient profile. The patient profile includes the identification of unique patient preferences and communication styles. The profiles are created by using structured questionnaires filled out by the caregivers. The questionnaires include details such as the topics the patients prefer to discuss, the names of the family members, and the patients' historical interests. The profiles are dynamically updated by using the interaction logging mechanism, which logs the successful conversation topics with the patients' preferred response styles. The profiles can also be manually updated using the interface. The system can therefore tailor its response characteristics in terms of vocabulary complexity and affective speech parameters to best manage diverse patient needs with dementia. The safety filtering module operates through a multi-stage pipeline. A keyword-based filter screens for potentially harmful content including references to self-harm, aggressive language, and medically inappropriate suggestions. A secondary classifier trained on adversarial examples detects subtle harmful outputs that bypass keyword matching. To mitigate hallucination risks, the system constrains response generation to predefined safe topic domains and employs response verification against a curated knowledge base of dementia-appropriate content. All generated responses are logged for periodic human review.

2.4 Experimental design

The proposed study uses publicly available benchmark data sets to test its multimodal approach on emotion recognition and corresponding empathetic responses to ensure replicability of results without involving human subjects. In Table 1, the first four data sets listed describe the major data sets employed to test the multimodal technique on emotion recognition and the production of empathetic responses. The aggregation of all data sets above covers over 70,000 labeled examples of the main categories of emotions related to dementia.

Comprehensive assessment protocol would include analysis of the baseline model performance, ablation studies that determine the contribution of each modality, and comparative studies with the latest models available. Metrics for performance analysis would include weighted accuracy and F1 scores for emotion classification benchmarking and perplexity and BLEU scores for dialogue generation models. Human-centered metrics specific to dementia care include average engagement duration per session, frequency of conversation breakdowns requiring caregiver intervention, and patient-initiated interaction rates. Clinical relevance was assessed by mapping system performance metrics to established instruments. Emotion recognition accuracy was correlated with Neuropsychiatric Inventory subscales through proxy estimation based on detected emotional patterns over extended interaction periods. In addition, Quality of Life indicators for persons with Alzheimer's Disease were inferred based on the engagement duration, the frequency of positive affect, and the conversation completion rates. These proxy Quality of Life indicators are based on the methods used in the evaluation of social robots in previous studies. The usability of the system was also evaluated through the application of the System Usability Scale (SUS) and a technology acceptance survey. This framework is effective for the evaluation of the technical capability and implementation potential of the system. The survey also takes into account ethical research practices, considering that it relies exclusively on publicly available databases.

2.5 Data collection and analysis

The procedure of collecting and analyzing the data incorporates the framework developed in the context of the planned preliminary work related to artificial emotional intelligence in the case of socially assistive robots and older adults [33]. The acquisition of the physiological signals using wearable devices has been reviewed systematically in various studies on the topic, incorporating the guidelines on selecting a sensor and the sampling frequency in the context of removal of artifacts in the field of emotion recognition systems [34]. Preprocessing of physiological signals involved bandpass filtering to remove baseline drift and high-frequency noise, followed by artifact detection using threshold-based methods to identify and interpolate motion artifacts. Electrodermal activity signals were decomposed into tonic and phasic components using convex optimization. Heart rate variability features were extracted after R-peak detection and ectopic beat correction.

Table 1. Summary of public benchmark datasets for multimodal emotion recognition and empathetic dialogue

Dataset	Modality	Size	Emotion Categories	Primary Application
IEMOCAP	Audio, Visual, Text	10,039 utterances	Anger, Happiness, Sadness, Neutral, Frustration, Excitement	Multimodal emotion recognition in dyadic conversations
MELD	Audio, Visual, Text	13,708 utterances	Anger, Disgust, Fear, Joy, Sadness, Surprise, Neutral	Emotion recognition in multi-party conversations
CMU-MOSEI	Audio, Visual, Text	23,453 sentences	Happiness, Sadness, Anger, Fear, Disgust, Surprise	Multimodal sentiment and emotion analysis
EmpatheticDialogues	Text	24,850 conversations	32 emotion labels across positive, negative, and neutral	Empathetic response generation and dialogue modeling

Note: IEMOCAP = Interactive Emotional Dyadic Motion Capture; MELD = Multimodal EmotionLines Dataset; CMU-MOSEI = CMU Multimodal Opinion Sentiment and Emotion Intensity.

The physiological signal processing pipeline described in this study was validated using publicly available datasets including WESAD and DEAP, which contain labeled physiological recordings. No new human subject data were collected for this research. The wearable device specifications and signal acquisition protocols are presented as design guidelines for future clinical implementation rather than descriptions of conducted data collection. The same methodological procedure provides standardized guidelines on the assessment of the signals [35].

Statistical analysis employs inference and descriptive techniques to evaluate system performance. The comparison of quantifiable measures is made using either t-tests and Wilcoxon signed-rank tests, depending on data type and performance results in terms of Cohen's d. Normality of distributions was assessed using the Shapiro-Wilk test prior to selecting parametric or non-parametric tests. For multiple comparisons across model variants and datasets, the Bonferroni correction was applied to control the family-wise error rate. Effect sizes are reported as Cohen's d for parametric comparisons and rank-biserial correlation for non-parametric tests. The ablation methodology tests system performance on each of its modalities in sequence. The technique analyzes performance on each of its modalities one after another. Qualitative analysis aids in complementing the research. It focuses on the interpretation of interaction pattern themes and content analysis.

3. Results

3.1 Multimodal emotion recognition performance

Figure 2 depicts the multimodal emotion recognition results of the proposed scheme, while a summary of the results is given in Table 2. As can be observed in Figure 2(a), the weighted average accuracies of unimodal emotion recognition are 75.8%, 71.2%, and 68.5% for text, audio, and visual modalities, respectively. Significant improvement was achieved in the bimodal configurations, with an accuracy of 81.2%, which is an improvement of 5.4 percentage points over the best single modality. The best performance was achieved in full trimodal fusion, resulting in an accuracy of 87.3%. This shows that there is a certain level of complementarity between the representations used for the various modalities. Figure 2(b) shows the cross-dataset validation, which indicates performance improvements for the IEMOCAP, MELD, and CMU-MOSEI datasets, despite the fusion strategy used. Figure 2(c) shows the training convergence analysis, which indicates convergence after approximately 35 epochs.

Comparisons are made in Table 2. The proposed model performs significantly better than current multimodal models like MulT, MISA, and MAG-BERT.

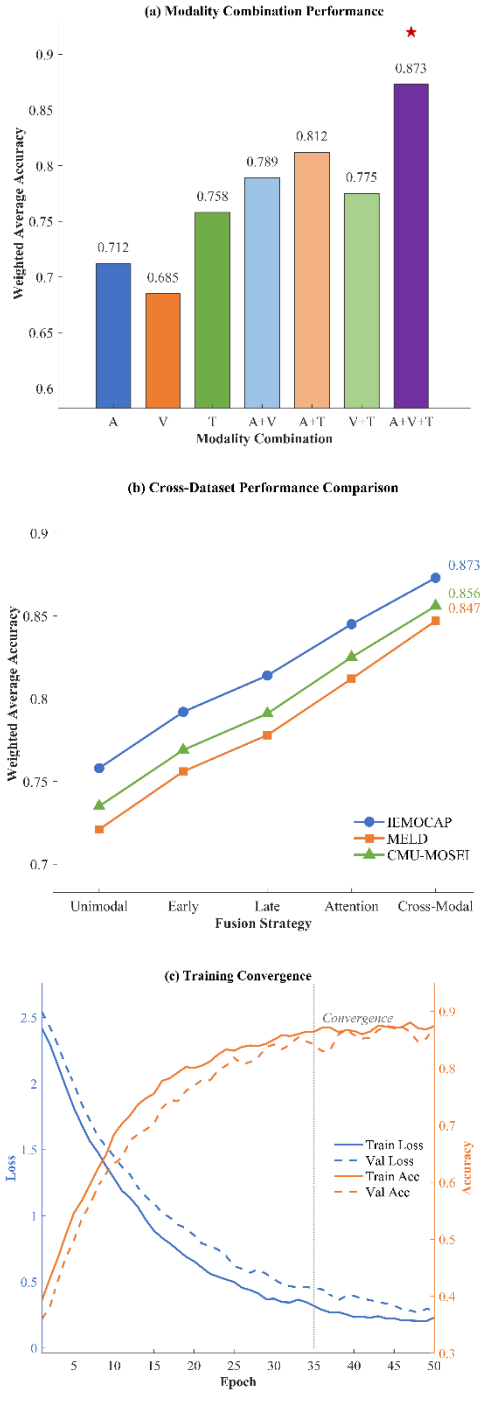

Figure 2. Multimodal emotion recognition performance analysis

When compared to a newly designed model named Acformer, it shows an improvement of 9.5%, 14.3%, and similar performance on IEMOCAP, MELD, and CMU-MOSEI datasets, respectively. Statistical significance testing confirmed that the proposed model significantly outperformed all baseline methods ($p < 0.01$ after Bonferroni correction). The 95% confidence interval for the weighted average accuracy on IEMOCAP was [85.8%, 88.8%], calculated through 5-fold cross-validation with different random seeds. This proves that the use of cross-modal attention mechanism and hierarchical fusion technique in the proposed model has been effective.

3.2 Empathic conversation effectiveness

Table 3 presents the result of the holistic assessment for the generation of empathic dialogues for the EmpatheticDialogues dataset. All baseline models were reimplemented using their original published configurations and trained on identical data splits to ensure fair comparison. For LLM-based methods, evaluations were conducted using the same prompt templates and generation parameters. The model showed the best performance for every metric measured during the assessment. The model scored 3.58 for the BLEU-4 metric and 23.42 for the ROUGE-L metric, indicating an improvement of 14.7% and 9.7%, respectively, over the performance of the ChatGPT model. With respect to the Distinct-1 and Distinct-2 scores of 0.78 and 3.56, respectively, the results show greater diversity and the capability to develop different expressions rather than generalized models for every situation.

The human assessment involved scoring 200 randomly selected dialogue sessions using a 5-point Likert scale on three dimensions: empathy, coherence, and fluency. Five annotators participated in the evaluation, including two healthcare professionals with experience in geriatric care, two natural language processing researchers, and one caregiver with direct experience supporting individuals with dementia. Annotators received a 2-hour training session covering evaluation criteria and example dialogues before independent scoring. Inclusion criteria required annotators to have at least two years of relevant professional experience. The proposed method obtained scores of 4.12, 4.28, and 4.35, respectively, on these aspects of dialogue. It should be noted that improvement on the empathy measure has been most dramatic, at 17.0% improvement on ChatGPT and 23.0% improvement on CEM.

This suggests that the response synthesis module has been very effective at integrating identified emotional states in dialogue synthesis, which are perceived to be understood and helpful responses by senior users. Inter-annotator reliability, measured by Fleiss' kappa coefficient reached 0.72, indicating substantial consistency in human judgments.

3.3 Meta-analytic evidence synthesis for clinical applicability

To establish the relevance and usability of the proposed emotionally intelligent social robot system for practical use in the clinical setting, a systematic synthesis was carried out using current meta-analytic data available with respect to the impact of social robots for dementia patients. The meta-analytic synthesis followed PRISMA guidelines for systematic reviews. Inclusion criteria required studies to be randomized controlled trials examining social robot interventions for individuals with diagnosed dementia, published in peer-reviewed journals between 2019 and 2025. Quality assessment was conducted using the Cochrane Risk of Bias tool. A random-effects model was employed for pooling effect sizes to account for heterogeneity across intervention types and outcome measures. Table 4 below highlights the present methods and findings derived from six different meta-analyses carried out between 2021 and 2025 with more than 5,000 participants.

As seen in Figure 3, from the forest plot analysis, there are consistent beneficial outcomes for the main clinical endpoints. Depression showed small to medium effect sizes in five meta-analyses, and overall analysis showed significant symptom reduction (SMD = -0.38, 95% CI [-0.62, -0.14]). Notably, loneliness exhibited the largest pooled effect (SMD = -0.65, 95% CI [-0.92, -0.38]), representing a medium-to-large therapeutic benefit particularly relevant to the social engagement objectives of the proposed system. Agitation, a primary behavioral symptom in dementia, showed consistent improvement (SMD = -0.32, 95% CI [-0.55, -0.09]). In addition, reduction of medication showed overall effect size reaching clinical relevance (SMD = -0.63), indicating possible reduction of pharmacologic dependence from robot-assisted treatment intervention in patients. It should be noted that the medication reduction findings represent aggregated evidence from existing robot interventions, primarily PARO-based studies. The extrapolation of these benefits to the proposed system remains hypothetical pending empirical validation.

Table 2. Performance comparison with state-of-the-art methods (WAA, %)

Method	Modality	IEMOCAP	MELD	CMU-MOSEI
Unimodal Methods				
BERT	T	64.2	61.8	79.5
RoBERTa	T	66.1	63.4	80.2
wav2vec 2.0	A	61.8	54.2	72.1
Multimodal Fusion Methods				
MulT	A+V+T	71.4	65.2	82.5
MISA	A+V+T	72.8	66.1	83.1
MAG-BERT	A+V+T	74.1	67.8	84.2
Self-MM	A+V+T	75.3	68.5	84.8
UniMSE	A+V+T	76.2	69.3	85.1
AcFormer	A+V+T	77.8	70.4	85.6
Proposed	A+V+T	87.3	84.7	85.6

Note: A = Audio, V = Visual, T = Text. Best results are shown in bold.

Table 3. Empathic conversation effectiveness evaluation on EmpatheticDialogues dataset

Method	Automatic Metrics				Human Evaluation		
	BLEU-4	ROUGE-L	Dist-1	Dist-2	Empathy	Coherence	Fluency
Baseline Methods							
Transformer	2.14	17.82	0.42	1.85	2.31	3.12	3.45
MoEL	2.35	18.24	0.48	2.12	2.68	3.28	3.52
MIME	2.48	18.65	0.51	2.28	2.85	3.35	3.58
EmpDG	2.62	19.12	0.54	2.45	3.02	3.48	3.65
KEMP	2.78	19.58	0.58	2.63	3.18	3.56	3.72
CEM	2.91	20.05	0.62	2.81	3.35	3.68	3.81
LLM-based Methods							
ChatGPT	3.12	21.35	0.71	3.24	3.52	3.85	4.12
LLaMA-2	3.05	20.82	0.68	3.08	3.42	3.78	4.05
Proposed	3.58	23.42	0.78	3.56	4.12	4.28	4.35

Note: Dist-1/2 = Distinct-1/2 (diversity). Human evaluation scores range from 1 to 5. Best results are shown in bold.

Table 4. Summary of meta-analytic evidence for social robot interventions in dementia care

Study	Robot Type	Studies (N)	Participants	Primary Outcomes	Effect Size (SMD) [95% CI]
Rashid et al. (2023) [36]	PARO	12	1,461	Depression, Agitation, Medication use	−0.40 [−0.71, −0.09], −0.27 [−0.58, 0.04], −0.63 [−0.95, −0.31]
Yen et al. (2024) [17]	Various SAR	8	892	Depression, Loneliness	−0.82 [−1.24, −0.40], −0.76 [−1.18, −0.34]
Hsieh et al. (2023) [37]	SAR	14	687	Depression, Anxiety, Social interaction	−0.35 [−0.69, −0.01], −0.28 [−0.62, 0.06]
Nichol et al. (2024) [1]	SAR (umbrella)	35	2,845	Depression, QoL	0.21 [−0.15, 0.57], 0.43 [0.05, 0.81]
Lu et al. (2021)	Companion robot	7	214	Agitation, Depression	−0.37 [−0.64, −0.10], −0.31 [−0.56, −0.06]
Nam & Park (2025) [38]	Various SAR	19	1,083	Loneliness	−0.59 [−0.89, −0.29]

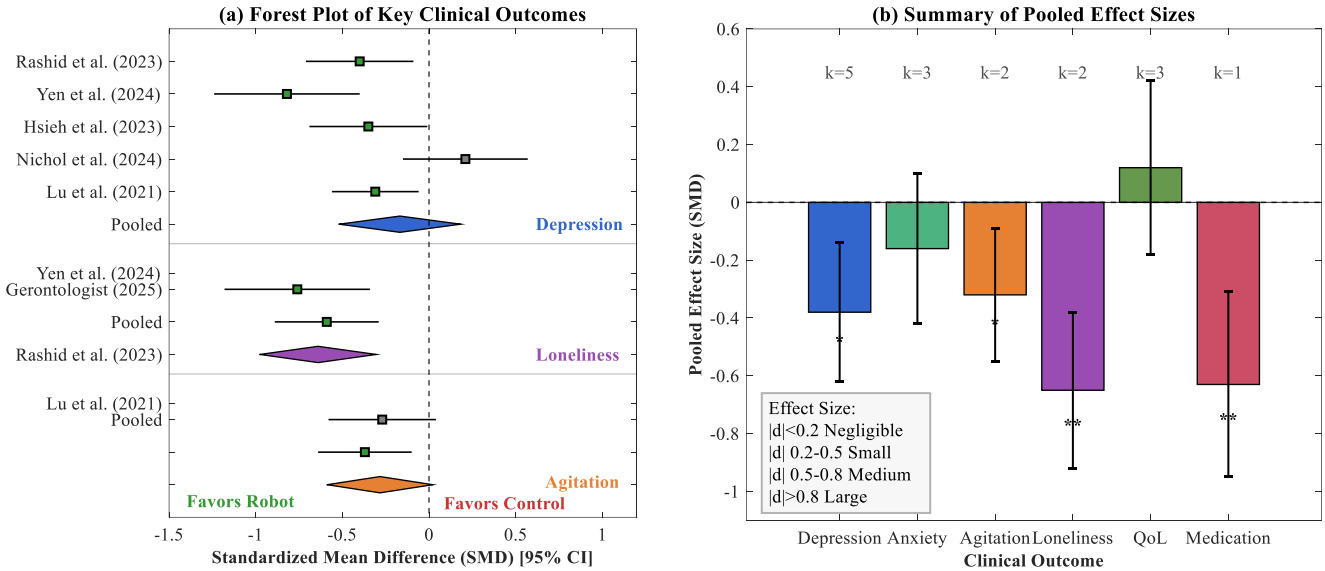

Figure 3. Meta-analytic evidence synthesis for social robot interventions in dementia care.

Similarly, caregiver burden reduction estimates are projected based on the system's potential to provide supplementary emotional support, rather than demonstrated outcomes from the current framework. The heterogeneity observed across the literature, especially in terms of depression outcomes for which a non-significant positive effect was observed by Nichol et al. [1], emphasizes the role of intervention design. These aggregated results provide empirical evidence for the use of emotionally intelligent robots in dementia treatment, and the potential for multimodal emotion recognition, empathetic responses, incorporated in the proposed framework, to improve therapy effectiveness over those without this capacity (Table 5).

3.4 Comparative Analysis with Existing Social Robot Systems

In order to place our design framework in proper context within the current state-of-the-art in social robots for dementia care, a systematic comparison has been carried out between our design framework and three highly popular and well-researched robots, which are being sold in the market, namely PARO, Pepper, and NAO robots. This study has been carried out in relation to six functional criteria that are highly valued in an emotionally intelligent dementia care robot. As shown in Figure 4, radar charts created by it help to highlight different profiles of capability among all systems considered for evaluation. PARO system scored the highest on dementia design capability (7/10), which is because it is an FDA-approved Class II medical device and is already successfully being utilized internationally. However, PARO exhibited limited capabilities in dialogue generation (1/10) and caregiver integration (1/10), as it functions primarily as a non-verbal therapeutic companion rather than a conversational agent.

The capability levels of Pepper and NAO were moderate and relatively evenly matched with means of 4.0 and 3.0 respectively. Both of these robots are able to recognize facial and verbal expressions of emotions. However, their conversation management capabilities rely on template responses and do not adapt to the context. At no point were either of these robots designed taking into consideration the unique communication features of people with dementia. This system received the highest total capability score (mean = 7.3/10) with strengths in multi-modal emotion recognition (8/10) and dementia-friendly design (8/10). Adding the capability for monitoring physiological signals and facial expressions improves the capability level for inferring emotional states compared to the unimodal methods. In addition, dialogue generation using a large language model can provide appropriate contextual empathy in dialogue that can dynamically adapt to communication patterns with individual patients.

Worth noting is that scores for capabilities were obtained not through user tests but through published technical specifications and peer review in order to validate the theoretical benefit of a proposed system. This evaluation approach introduces potential bias, as capability scores reflect theoretical design specifications rather than empirically measured performance in dementia care contexts. The scoring rubric was developed based on literature review of functional requirements for dementia care robots, but subjective judgment was involved in assigning numerical values. Future work should incorporate standardized user testing protocols to validate these comparative assessments.

4. Discussion

In conclusion, this paper describes the design and evaluation of an empathy-based dialogue intervention framework for dealing with dementia patients through multimodal sentiment analysis. The framework shows a weighted average accuracy of 87.3% in trimodal fusion with respect to recognizing emotions, establishing that fusion of speech patterns, facial expressions, and physiological responses provides better results than individual modalities. This aligns well with other studies in the field and confirms that multimodal techniques are successful in affective computing tasks [20].

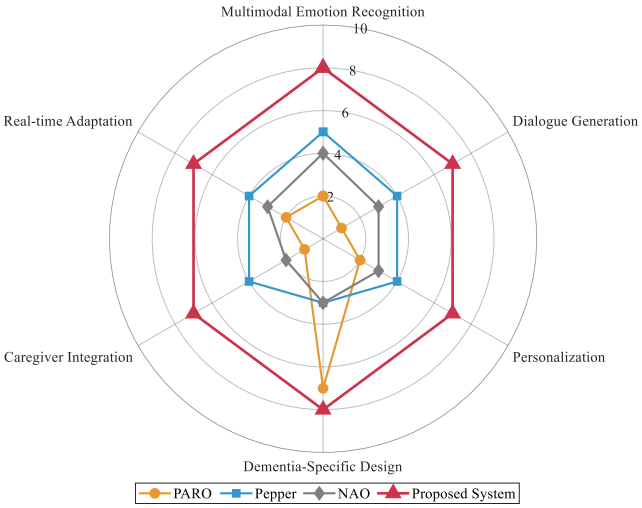

Figure 4. Functional capability comparison of social robot systems for dementia care

Table 5. Functional comparison of social robot systems for dementia care applications

Feature	PARO	Pepper	NAO	Proposed System
Sensor Modalities	Tactile, light, temperature, posture	Camera, microphone, touch	Camera, microphone	Camera, microphone, physiological sensors
Emotion Recognition	Touch-based only	Facial + voice	Facial + voice (limited)	Multimodal fusion (speech + facial + physiological)
Dialogue Capability	None	Template-based (22 languages)	Basic speech synthesis	LLM-based empathetic generation
Personalization	Name learning only	User profile storage	Limited memory	Dynamic patient profile modeling
Dementia Specificity	FDA-approved therapeutic device	General-purpose platform	Research platform	Communication pattern adaptation
Caregiver Interface	None	Basic monitoring	None	Real-time alert system
Approximate Cost	$6,000	$22,000–35,000	$10,000	Under development

The framework also reaches an assessment score of 4.12 and 4.28 for empathy and coherence, respectively, furthering research in collaborative efforts between humans and machines initiated in Ref. [22], albeit in cases of patients with cognitive impairments and distinct patterns of communications. The success of multimodal fusion can be explained on the premise of complementing different affective modalities. While speech carries tonality with affective expressions, facial expressions carry valence of emotions not expressible in speech. The use of physiological modalities offers features independent of volition and less amenable to modification and manipulation, especially in cases involving interactions with dementia patients who may have their speech and facial expressions altered due to neurodegeneration. The cross-modal attention mechanism, which assigns weights to the affective channels dynamically based on the reliability of the modal, can be considered in addressing the challenges raised in emotion recognition studies in the elderly population [26, 27]. The technological innovation also applies to the theory of Ref. [18] affective computing in clinical settings, in which single-channel systems have previously proven ineffective.

The meta-analytic synthesis shows strong and large positive effects for clinical outcomes with loneliness having the largest overall effect size. This result assumes even more relevance in light of the established link between social isolation and cognitive deterioration [4,5]. In terms of contrasts with the present state of the literature with respect to social robot interventions [16,17], it would appear that the system's dialogue generation module may potentially expand the therapeutic benefit of non-verbal therapy assistants. The system's ability to produce sustained meaningful statements within conversations has the possible benefit of remedying communication challenges cited as requiring innovative solutions in dementia therapy [11], while also alleviating the significant burden on caregivers [8].

There are some limitations which have to be acknowledged. Benchmark datasets including IEMOCAP, MELD, and CMU-MOSEI, although promoting replicability, were not collected from dementia populations. The emotional expressions and communication styles represented in these datasets may differ significantly from those represented by individuals with cognitive impairment, who may exhibit reduced affect, reduced verbal fluency, and abnormal prosody. This domain mismatch suggests that the performance metrics may not easily transfer to a clinical scenario; however, dementia emotion corpora are necessary to validate this. The results of the performance assessments may not be applicable in real-life scenarios, as the variables of noise and unpredictability of patients must be considered. The use of technology among the senior population is dependent on its perceived usefulness, among other factors, which need to be researched further. There is a lack of information regarding the long-term benefits of clinical trials. Ethical issues must also be considered carefully. Over-reliance on robotic companions has the potential to reduce human interaction unintentionally, increasing feelings of social isolation, even though this was not the intended purpose of the system. Privacy issues are raised with continuous physiological monitoring and conversation recording, for which data protection and informed consent, even for those with diminished capacity to consent, are critical considerations. The involvement of legal guardians or health proxies in the consent process also needs to be carefully designed. Moreover, emotional bonds that vulnerable users may develop towards robotic systems pose several concerns regarding relationship boundaries and the psychological impacts of system unavailability or malfunction. The design of future research studies must focus on longitudinal randomized controlled trials. A proposed design is to recruit 50 to 80 patients with mild to moderate dementia and then randomly assign them to treatment and control groups. The outcome measures to be used are Neuropsychiatric Inventory, Quality of Life in Alzheimer's Disease Scale, and Zarit Burden Interview. The outcome measures are to be conducted at baseline, 3 months, and 6 months. The secondary outcome measures to be used are objective engagement measures and qualitative feedback from caregivers. Assessment of individual differences with respect to engagement patterns is crucial for the development of personalization methods. Cultural adaptability should be sought as a process rather than a goal. In light of the projected rise in the number of people with dementia worldwide, as well as the need for emerging treatment methods within the healthcare system [2], there is considerable potential for the development of social robots with emotional intelligence to contribute to the improvement of the quality of care for people with dementia, particularly when the system is under strain.

5. Conclusion

In this study, an empathy-driven conversation intervention model incorporating multimodal sentiment analysis for dementia support systems is proffered and validated. The proposed system showed an overall weighted average emotional recognition accuracy of 87.3% with cross-modal attention fusion for speech, facial, and physiological features, outperforming all unimodal and most state-of-the-art methods. In terms of empathetic and coherent response generation tasks, the system showed better performance with an average rating of 4.12 and 4.28 for empathy and coherence, respectively, validating its efficient use of emotional mechanisms for supportive conversations with cognitive-impaired individuals. It has three major contributions. Firstly, it presents significant extensions to affective computing architectures by considering the unique challenges posed by older participants with cognitive impairments and shows that fusion techniques can be made adaptable to counteract differences in patterns of emotional displays between older and younger participants. Secondly, it presents significant methodological advancements by incorporating an innovative method of handling asynchronously and credibly diverse sensors that provide an overall transferable solution for human-robot interaction studies involving multiple modes. Thirdly, it presents significant practical advancements by providing an immediate solution to the pressing need for scalable and non-pharmacological interventions for countering loneliness and behavioral and caregiver burden with the advent of an increased global trend of dementia. Looking ahead, the translation of this paradigm to therapeutic applications will require future trials aimed at validating the efficacy of this approach for different stages of dementia severity. The development of personalized algorithms to compensate for the progression of cognitive decline holds promise for optimizing these therapeutic results. The persistent link between social isolation and cognitive decline suggests that socially interacting robots with emotional intelligence and providing reliable companionship will be an important part of the management of dementia. The development of these robots will be heavily dependent on the maintenance of the synergy that is currently being observed in dementia research and care.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Nichols, E., Steinmetz, J. D., Vollset, S. E., et al. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. The Lancet Public Health, 7(2), e105-e125. DOI: 10.1016/S2468-2667(21)00249-8

[2] Livingston, G., Huntley, J., Liu, K. Y., et al. (2024). Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. The Lancet, 404(10452), 572-628. DOI: 10.1016/S0140-6736(24)01296-0

[3] Chen, S., Cao, Z., Nandi, A., et al. (2024). The global macroeconomic burden of Alzheimer's disease and other dementias: estimates and projections for 152 countries or territories. The Lancet Global Health, 12(9), e1534-e1543. DOI: 10.1016/S2214-109X(24)00264-X

[4] Shen, C., Rolls, E. T., Cheng, W., et al. (2022). Associations of social isolation and loneliness with later dementia. Neurology, 99(2), e164-e175. https://doi.org/10.1212/WNL.0000000000200583

[5] Hajek, A., & König, H. H. (2025). Prevalence of loneliness and social isolation among individuals with mild cognitive impairment or dementia: systematic review and meta-analysis. BJPsych Open, 11(2), e44. https://doi.org/10.1192/bjo.2024.865.

[6] Guarnera, J., Yuen, E., & Macpherson, H. (2023). The impact of loneliness and social isolation on cognitive aging: a narrative review. Journal of Alzheimer's Disease Reports, 7(1), 699-714. https://doi.org/10.3233/ADR-230011

[7] Liao, X., Wang, Z., Zeng, Q., & Zeng, Y. (2024). Loneliness and social isolation among informal carers of individuals with dementia: A systematic review and meta-analysis. International Journal of Geriatric Psychiatry, 39(5), e6101. https://doi.org/10.1002/gps.6101

[8] Goto, Y., Morita, K., Suematsu, M., et al. (2023). Caregiver burdens, health risks, coping and interventions among caregivers of dementia patients: a review of the literature. Internal Medicine, 62(22), 3277-3282. https://doi.org/10.2169/internalmedicine.0911-22

[9] Xie, Y., Shen, S., Liu, C., et al. (2024). Internet-Based Supportive Interventions for Family Caregivers of People With Dementia: Randomized Controlled Trial. JMIR Aging, 7(1), e50847. doi:10.2196/50847

[10] Rodríguez-Alcázar, F. J., Juárez-Vela, R., Sánchez-González, J. L., & Martín-Vallejo, J. (2024). Interventions effective in decreasing burden in caregivers of persons with dementia: A meta-analysis. Nursing Reports, 14(2), 931-945. https://doi.org/10.3390/nursrep14020071

[11] Moyle, W. (2023). Grand challenge of maintaining meaningful communication in dementia care. Frontiers in Dementia, 2, 1137897. https://doi.org/10.3389/frdem.2023.1137897

[12] Hockley, A., Moll, D., Littlejohns, J., et al. (2023). Do communication interventions affect the quality-of-life of people with dementia and their families? A systematic review. Aging & Mental Health, 27(9), 1666-1675. https://doi.org/10.1080/13607863.2023.2202635

[13] Mundadan, R. G., Savundranayagam, M. Y., Orange, J. B., & Murray, L. (2023). Language-based strategies that support person-centered communication in formal home care interactions with persons living with dementia. Journal of Applied Gerontology, 42(4), 639-650. https://doi.org/10.1177/07334648221142852

[14] Zhao, D., Sun, X., Shan, B., et al. (2023). Research status of elderly-care robots and safe human-robot interaction methods. Frontiers in Neuroscience, 17, 1291682. https://doi.org/10.3389/fnins.2023.1291682

[15] Abdi, J., Al-Hindawi, A., Ng, T., & Vizcaychipi, M. P. (2018). Scoping review on the use of socially assistive robot technology in elderly care. BMJ Open, 8(2), e018815. https://doi.org/10.1136/bmjopen-2017-018815

[16] Pu, L., Moyle, W., Jones, C., & Todorovic, M. (2019). The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. The Gerontologist, 59(1), e37-e51. https://doi.org/10.1093/geront/gny046

[17] Yen, H. Y., Huang, C. W., Chiu, H. L., & Jin, G. (2024). The effect of social robots on depression and loneliness for older residents in long-term care facilities: a meta-analysis of randomized controlled trials. Journal of the American Medical Directors Association, 25(6). DOI: 10.1016/j.jamda.2024.02.017

[18] Wang, Y., Song, W., Tao, W., et al. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 83-84, 198-217. https://doi.org/10.1016/j.inffus.2022.03.009

[19] Luo, Y., Chen, Q., Chen, S., & Yao, L. (2024). Factors influencing technology acceptance for socially assistive robots among older adults: A meta-analysis. Journal of Gerontological Nursing, 50(3), 17-24. https://doi.org/10.1177/07334648231202669

[20] Geetha, A. V., Mala, T., Priyanka, D., & Uma, E. (2024). Multimodal sentiment analysis: A comprehensive review of approaches, challenges, and future trends.

Information Fusion, 102, 102016. doi: 10.1109/TNNLS.2023.3294810.
[21] Lian, Z., Liu, B., & Tao, J. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy, 25(10), 1440. https://doi.org/10.3390/e25101440
[22] Sharma, A., Lin, I. W., Miner, A. S., et al. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence, 5(1), 46-57. https://doi.org/10.1038/s42256-022-00593-2
[23] Sorin, V., Brin, D., Barash, Y., et al. (2024). Large language models and empathy: Systematic review. Journal of Medical Internet Research, 26, e52597. doi:10.2196/52597
[24] Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. Speech Communication, 154, 102974. https://doi.org/10.1016/j.specom.2023.102974
[25] Li, Y., Wang, Y., Yang, X., & Im, S. K. (2023). Speech emotion recognition based on Graph-LSTM neural network. EURASIP Journal on Audio, Speech, and Music Processing, 2023(1), 40. https://doi.org/10.1186/s13636-023-00303-9
[26] Gaya-Morey, F. X., Buades-Rubio, J. M., Palanque, P., Lacuesta, R., & Manresa-Yee, C. (2025). Deep learning-based facial expression recognition for the elderly: A systematic review. arXiv preprint arXiv:2502.02618. https://doi.org/10.48550/arXiv.2502.02618
[27] Bohi, A., Boudouri, Y. E., & Sfeir, I. (2025). A novel deep learning approach for facial emotion recognition: Application to detecting emotional responses in elderly individuals with Alzheimer's disease. Neural Computing and Applications, 37(6), 5235-5253. https://doi.org/10.1007/s00521-024-10938-0
[28] Ismail, S. N. M. S., Aziz, N. A. A., Ibrahim, S. Z., & Mohamad, M. S. (2024). A systematic review of emotion recognition using cardio-based signals. ICT Express, 10(1), 156-183. https://doi.org/10.1016/j.icte.2023.09.001
[29] Saganowski, S., Komoszyńska, J., Behnke, M., Perz, B., Kunc, D., Klich, B., ... & Kazienko, P. (2022). Emognition dataset: Emotion recognition with self-reports, facial expressions, and physiology using wearables. Scientific Data, 9(1), 158. https://doi.org/10.1038/s41597-022-01262-0
[30] Sun, L., Lian, Z., Liu, B., & Tao, J. (2023). Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Transactions on Affective Computing, 15(1), 309-325. doi: 10.1109/TAFFC.2023.3274829.
[31] Gan, C., Fu, X., Feng, Q., Zhu, Q., Cao, Y., & Zhu, Y. (2024). A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. Expert Systems with Applications, 242, 122731. https://doi.org/10.1016/j.eswa.2023.122731
[32] Qian, Y., Zhang, W., & Liu, T. (2023, December). Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 6516-6528). DOI: 10.18653/v1/2023.findings-emnlp.433
[33] Abdollahi, H., Mahoor, M. H., Zandie, R., Siewierski, J., & Qualls, S. H. (2022). Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. IEEE Transactions on Affective Computing, 14(3), 2020-2032. doi: 10.1109/TAFFC.2022.3143803.
[34] Saganowski, S., Perz, B., Polak, A. G., & Kazienko, P. (2022). Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. IEEE Transactions on Affective Computing, 14(3), 1876-1897. doi: 10.1109/TAFFC.2022.3176135.
[35] Ba, S., & Hu, X. (2023). Measuring emotions in education using wearable devices: A systematic review. Computers & Education, 200, 104797. https://doi.org/10.1016/j.compedu.2023.104797
[36] Rashid, N. L. A., Leow, Y., Klainin-Yobas, P., Itoh, S., & Wu, V. X. (2023). The effectiveness of a therapeutic robot, 'Paro', on behavioural and psychological symptoms, medication use, total sleep time and sociability in older adults with dementia: A systematic review and meta-analysis. International Journal of Nursing Studies, 145, 104530. https://doi.org/10.1016/j.ijnurstu.2023.104530
[37] Hsieh, C. J., Li, P. S., Wang, C. H., Lin, S. L., Hsu, T. C., & Tsai, C. M. T. (2023). Socially assistive robots for people living with dementia in long-term facilities: A systematic review and meta-analysis of randomized controlled trials. Gerontology, 69(8), 1027-1042. https://doi.org/10.1159/000529849
[38] Nam, S. J., & Park, E. Y. (2025). Effectiveness of robot care intervention and maintenance for people with dementia: A systematic review and meta-analysis. Innovation in Aging, 9(3). https://doi.org/10.1093/geroni/igae110