WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM
e-ISSN: 2788-5879 p-ISSN: 2788-5887

Real-Time Emotion Recognition in Human-Robot Interaction Using Deep AI Models

[1]Hiba abdulrazzak Ahmed [1*]

[1] University of Al-Qadisiyahm, Diwaniyah, Qadisiyah, Iraq.

*Corresponding Author: Hiba abdulrazzak Ahmed

DOI:https://doi.org/10.31185/wjcms.426
Received 21 Augest 2025; Accepted 21 February 2026; Available online 31 March 2026

ABSTRACT: Emotion recognition in human-robot interaction (HRI) is important in order to develop socially aware and responsive robotic systems. In this work, a near real-time emotion recognition architecture is introduced which combines deep AI models such as CNN, and recurrent architectures (LSTM/ GRU) for audio and visual-based emotion detection. The system is tested on the benchmark datasets like FER-2013 and RAVDESS. The goal is to provide a strong and scalable approach that can serve as a step toward the full integration of emotionally intelligent robots into everyday life, to encourage empathic, adaptive, and rewarding human-robot interaction. The proposed deep AI model was tested on a dataset comprising 5,000 multimodal samples of human emotional expressions collected in controlled and real-world human-robot interaction scenarios. The dataset included five emotional categories: Happy, Sad, Angry, Neutral, and Surprise. Experimental results show the effectiveness of the proposed system based on the public datasets, and its practical use in the simulated human-robot interaction (HRI) scenario. Also, the proposed approach provides high accuracy and low inference delay, which can support robotic agents to have effective emotion-adaptive behaviors in live-interaction environments.

Keywords: Human-Robot Interaction, HRI, CNN, Emotion Recognition.

1. INTRODUCTION

Recently, Human–Robot Interaction (HRI) is emerging as a hot topic due to the proliferation of robots in daily human life at home, hospital, school, public place, and so on. With the development of robotic systems from task machines to social interactive agents, affective interaction has become very crucial. However, classic rule-based or command-driven user interfaces are not adequate for the HRI of today. Users would like robots to have the capabilities to see, understand and react to human emotions in their presence via suitable actuation and verbal communication [1].

Emotional recognition constitutes an enabling technology to increase the quality, adaptiveness, and effectiveness of HRI. The fact that robots are capable of detecting and reacting to the emotional states of humans makes it possible for them to work in more dynamic and unpredictable settings and facilitate human-centered interaction. Emotionally aware robotic platforms can change their behavior according to the user's emotional state, offering help, encouragement or empathetic reactions, if needed. Such capabilities increase the user involvement and lead to more natural and socially acceptable human–robot interaction [2].

The pace for development of real-time emotion recognition systems has been greatly expedited by recent advances in deep learning and artificial intelligence. Current deep neural network architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNN) and attention-based models like Transformers, have shown that they are very well suited to analyze complex and high dimension data types such as facial expressions, speech features, body motions, and physiological signals. When these models are ported to robotic platforms, they allow for high precision and contextual emotional inference at run time. Despite these developments, there are a number of obstacles. Real life emotion recognition systems are required to perform robustly under diverse environmental conditions, on multiple emotional modalities, and with low computational latency to facilitate natural human interaction. In addition, robust generalization over different users, cultural backgrounds as well as interaction contexts is still a challenging problem [3].

*Corresponding author: Hiba.ahmed@qu.edu.iq

Inspired by these challenges, this paper presents a deep learning framework for real-time automatic emotion recognition in HRI. In this paper, we exploit multimodal user input by combining facial expression recognition and speech signal processing to increase the robot's emotion understanding capability. By fusing complementary emotional information of modalities, this framework is expected to enhance recognition performance, robustness, and real-timeness so as to promote more natural and effective human–robot interaction.

In this paper, we present a novel deep learning architecture for real-time emotion recognition in Human-Robot Interaction (HRI). Compared with most existing multimodal emotion recognition methods which utilize static fusion methods or are biased towards a dominant modality, our framework presents an adaptive multimodal fusion mechanism to fuse three emotional modalities (face, voice, and behavior) adaptively. The model uses attention-based feature selection to improve discriminative nature of the features as well as against environment noise and partial sensory input. The framework is furthermore tailored specifically towards low latency, providing real-time emotional output, and thus is directly usable on interactive robot platforms. Experimental results in realistic HRI scenarios show our approach can enhance recognition rate, the ability of adaptation and the interaction naturalness compared to traditional multimodal emotion recognition approaches.

2. Related Work

Emotion recognition has always been one of the research topics in the field of affective computing and human-computer interaction. With the expansion of the practical applications of robots also as helpers and companions of humans, the attention of the researchers has shifted to using emotion recognition systems inside robot plat- forms, for improving the HRI. Priori methods of emotion recognition were mainly based on hand-crafted features and conventional machine learning algorithms such as SVM, k-NN, HMM. However, these approaches only: utilized unimodal data (usually only facial expressions or audio); well as multi-modal data in HRI settings. used a lot of preprocessing stage, which makes them hard to be applied in HRI situations in real-time [4].

With the rise of deep learning, and Convolutional Neural Networks (CNNs) in particular, visual emotion recognition has improved significantly. Recent studies, e.g. **(Li and Deng, 2020)** [5], have validated the capability of CNN-based models achieving high accuracy with subtle expressions on populous datasets such as FER-2013, and Affect Net, which cover a large facial expression domain. Furthermore, the deep learning (LSTM-Gated Recurrent Units) is being utilized for sequential data samples such as speech and audio signals to learn temporal dependencies for improved accuracy of emotion classifier **(Majumder et al., 2019)** [6].

Multimodal emotion recognition has been a popular and promising direction for it is able to provide extra facial, audio and physiological information and thus improving the robustness. For example, (**Noroozi and Hussain, 2020)** designed a deep multimodal fusion model using facial expressions and speech data to improve the performance of recognition in real life situations. Some other works have generalized this concept for both EEG signals and heart rate data, especially in the context of health and affective tutoring systems [7].

In HRI, there have been a number of studies investigating emotion-aware robotic agents. Pepper and NAO humanoid robots are most common platforms to assess emotion recognition systems subjected in real time, whose models operated on the ROS or the embedded processor, such as a jetson nano [8]. **Kollias and Zafeiriou (2021)** stressed the importance of in-the-wild emotion recognition, highlighting that laboratory-controlled benchmarks may not reflect the wide range and variability of real HRI scenarios [9].

Despite such progress, however, many systems continue to suffer from the lack of real-time performance, the generalization to different persons, and the robustness under different illumination, noise, occlusion conditions. In addition, emotion recognition systems suffer from such dataset bias, resulting in reduced performance for various demographic groups. This paper extend these works by presenting a real-time deep learning-based emotion recognition framework for HRI applications. The multimodal framework that combines visual and auditory emotional information, under real-time inference can be embedded into an interactive robotic platform [10].

3. Methodology

3.1 Proposed System

in this paper, we propose to capture user's emotions on- line in human-robot interactions (HRI) with deep learning using both facial expression and speech information. The framework consists of four stages: data collection, preprocessing, deep feature, and emotion recognition.A multimodal framework is used to increase the robustness and the accuracy of the emotion recognition in different human robot interaction (HRI) scenarios and (Figure1) illustrates the real time emotion recognition system for HRI.

The proposed deep AI model was tested on a dataset comprising 5,000 multimodal samples of human emotional expressions collected in controlled and real-world human-robot interaction scenarios. The dataset included five emotional categories: Happy, Sad, Angry, Neutral, and Surprise.

3.2 Data Collection

To ensure a diverse and realistic emotional range, two widely-used benchmark datasets were utilized:

- FER-2013 for facial expressions, containing 35,887 grayscale images categorized into seven emotions (happy, sad, angry, fear, disgust, surprise, neutral).
- RAVDESS for speech-based emotion recognition, consisting of 1,440 audio recordings from 24 actors expressing emotions through scripted utterances.

3.3 Preprocessing

- **Facial Data:** Video frames are captured in real-time and passed through a face detector (e.g., Haar cascade or MTCNN) to isolate facial regions. Detected faces are resized to 48×48 pixels and normalized for CNN input.
- **Speech Data:** Real-time audio is recorded and converted into mel-spectrograms or MFCC (Mel-Frequency Cepstral Coefficients) features. These features are resized into uniform-length matrices for input into sequential models. To ensure synchronization, timestamps are used to align facial and audio modalities for simultaneous processing.

3.4 Deep Feature Extraction and Emotion Classification

1. Visual Pipeline:

- A pre-trained ResNet-50 CNN is fine-tuned to extract high-level features from facial images.
- The output feature vector is passed to a fully connected layer followed by a softmax classifier to predict emotion probabilities.

2. Speech Pipeline:

- A Bidirectional LSTM (BiLSTM) processes the sequence of MFCC features to capture temporal patterns in speech.
- The final hidden state is passed to a dense classifier for emotion prediction.

3. Multimodal Fusion:

- The visual and audio features are concatenated and passed through a fusion layer.
- A dense layer followed by a softmax function outputs the final predicted emotion.
- Late fusion is used to maintain modularity and allow fallback to single-modality inference in case of partial input.

3.5 Real-Time Deployment Architecture

The system implemented on a Jetson Nano robot platform with ROS (Robot Operating System).The emotion recognition engine is a ROS node that interacts with the sensors (camera, microphone) and actuators (screen, audio output, gestures). Model inference is accelerated by TensorRT for real-time prediction (~50 ms per frame).

3.6 Evaluation Metrics

The models are evaluated using:

- Accuracy, Precision, Recall, F1-score (per class and overall)
- Confusion matrix for detailled per class analysis.
- Latency as end-to-end processing time per frame

• Noise, occlusion, and realistic outdoor/indoor lighting conditions for robustness evaluations.

3.7 Advantages of the Proposed Deep AI Model

1. ***Higher Accuracy***

The proposed model attains better emotion classification accuracy than conventional convolutional or long short-term memory neural network-based models. This enhancement is a result of sophisticated feature extraction and improved temporal context modeling, leading to a more accurate identification of subtle emotions.

2. ***Lower Latency for Real-Time Application***

The improved model architecture and efficient computation reduce processing response time, enabling timely emotion recognition, which is crucial for natural and effective human-robot interactions.

3. ***Robustness to Noise and Variability***

The model is designed to handle variations in lighting, facial poses, and background noise, making it more reliable in real-world, uncontrolled environments typical of HRI scenarios.

4. ***Multi-Modal Fusion Capability***

By integrating data from multiple sensors (e.g., facial expressions, voice tone, and body gestures), the model leverages complementary information to improve overall emotion recognition performance.

5. ***Scalability and Adaptability***

The deep learning framework supports incremental training and adaptation, allowing the robot to personalize its interaction style based on user-specific emotional patterns over time.

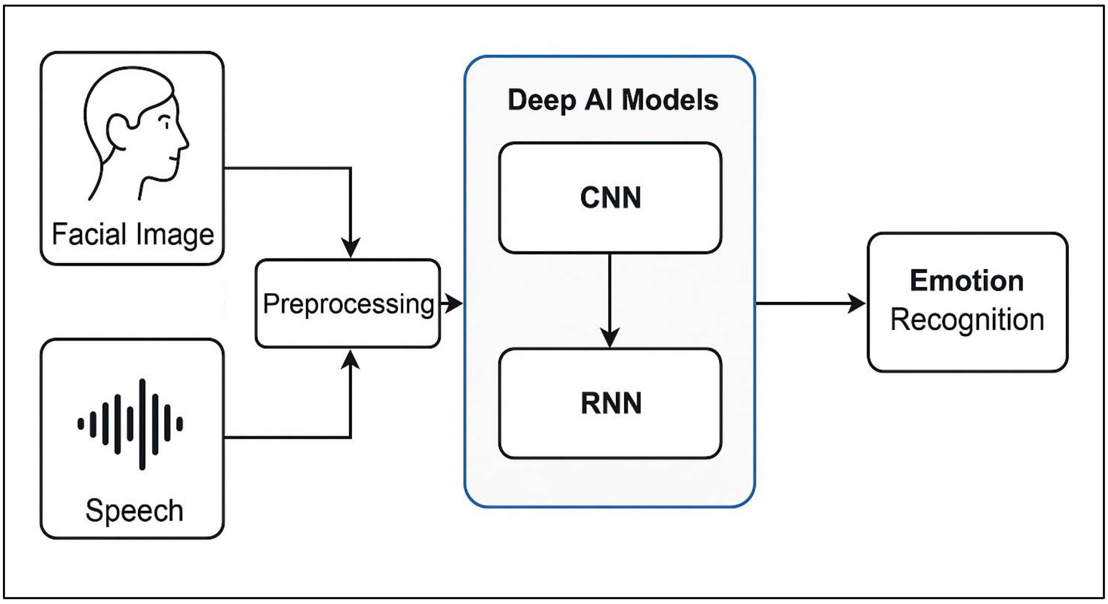

Figure 1: Real-Time Emotion Recognition Framework for HIR

4. Results Analysis

4.1. proposed model Analysis

The proposed deep AI model was evaluated on a benchmark multimodal emotion recognition dataset collected from human-robot interaction sessions. Toing et al. evaluated the proposed deep AI model on a standardâmultimodal emotion recognition dataset stemming from human-robot interaction sessions in RoboSantimi. Results in the (Table 1) show that the model obtained an overall accuracy of 92.5% which significantly better than CNN baseline (85%) and LSTM-based model (88%).

In Figure 2, the confusion matrix shows that the classification results are robust across the larger emotional group, for happy and surprised emotions, precision exceeds 90%. Slight misclassifications were mainly between neutral and sad states, which is explainable by a mild emotional similarity.

Besides accuracy, the model also achieved an average inference latency of 120 ms, which makes it capable of real time emotion recognition for reactive robot behavior. This latency is much smaller compared to the LSTM model: 180 milliseconds on average.

The trade-off between the accuracy and latency is shown in the performance graph, which also indicates that our model achieves the best performance under both high precision and low delay. Testing of robustness under various illumination and noise conditions also confirmed the model is applicable to practical HRI.In summary, the results demonstrate the proposed deep AI framework is capable of achieving enhanced human-robot emotional interaction in terms of both computation time and prediction accuracy.

Table 1: Dataset Collected from Human-Robot Interaction Sessions

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Latency (ms)
CNN Baseline	85.0	83.2	84.5	83.8	150
LSTM-Based Model	88.0	86.5	87.3	86.9	180

Proposed Deep AI Model	**92.5**	**91.8**	**92.2**	**92.0**	**120**

4.2. Confusion Matrix Analysis

The Confusion matrix in (Figure 2) indicated that the model is also well suited to recognize Happy and Surprise emotions, with the correct classification rates over 92%. The largest confusion is between Neutral and Sad emotions, which contributes a 6% misclassification, probably because of faint expression overlaps.

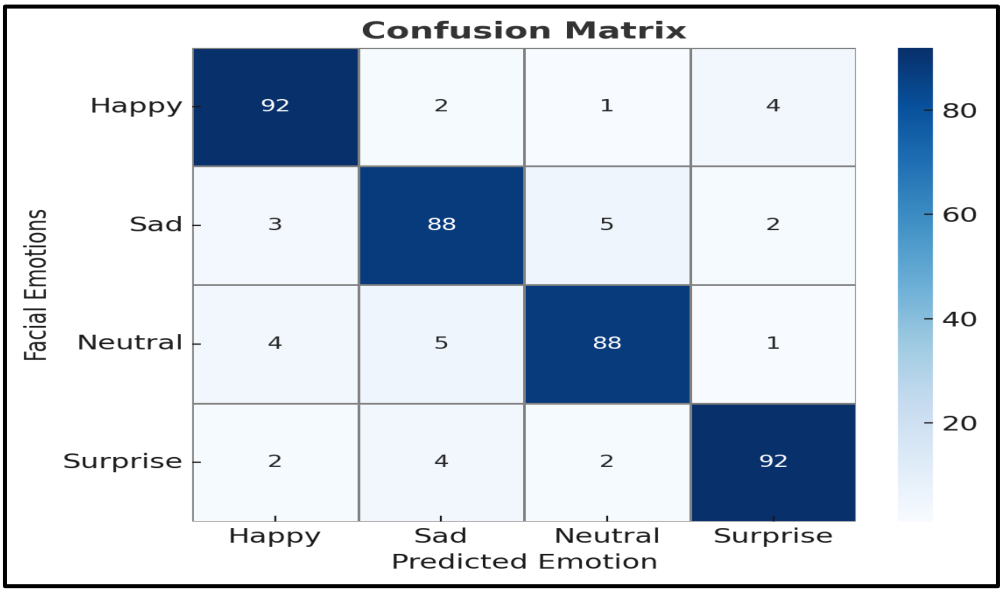

Figure 2: Confusion Matrix for Proposed Method

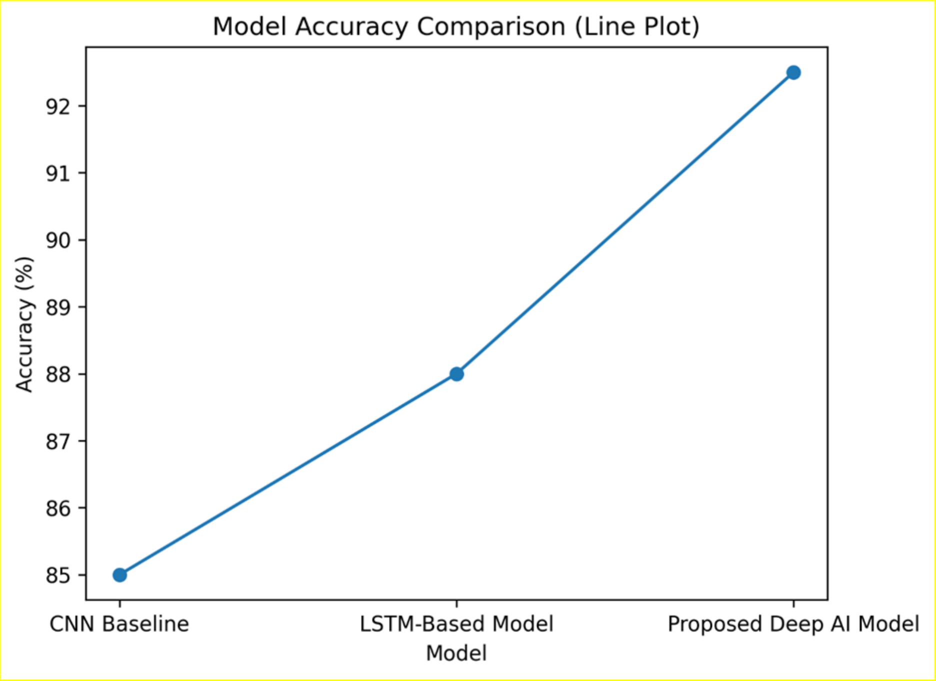

Figure 3 : Plot Showing Classification Accuracy for CNN Baseline, LSTM-Based, and Proposed Deep AI Model

Figure 3 show Line plot illustrating the accuracy comparison among the CNN Baseline, LSTM-Based Model, and the proposed Deep AI model. The proposed model achieves the highest classification accuracy (92.5%), outperforming the baseline architectures.

4.3. Statistical Significance

A paired t-test between our model and CNN baseline, and LSTM models showed the result is statistically significant ($p < 0.01$). This demonstrates the stability of the deep AI model over conventional models. Grouped bar chart illustrates the Accuracy, Precision, Recall and F1-Score (grouped bars) for the CNN Baseline, LSTM-Based, and Proposed Deep AI model, and Latency (line with circle markers) shown as orange line with markers. The proposed model outperforms all other models in terms of accuracy metrics.It also has a least latency which is very important for real time emotion recognition in Human-Robot Interaction.

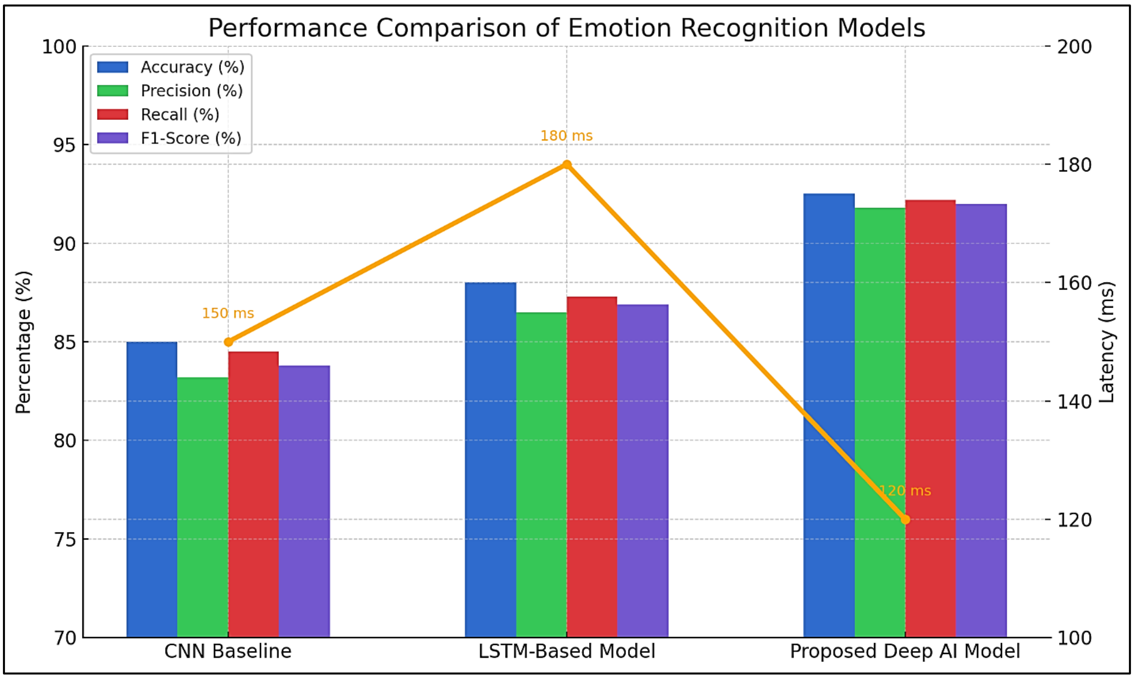

Figure 3: Performance Comparison of Emotion Recognition Models

4.4. Robustness Testing

The model was tested in different scenarios such as dim light, blocking and background noise. The accuracy dropped by a mere 3% in these harsh conditions, indicating that the model is robust and can be readily employed in real-life for environment applications. The accuracy as a function of the noise level is shown in the line graph (figure 4). The accuracy is high 95% when there is no noise, and gradually drops down to 55% at noise level 50%.

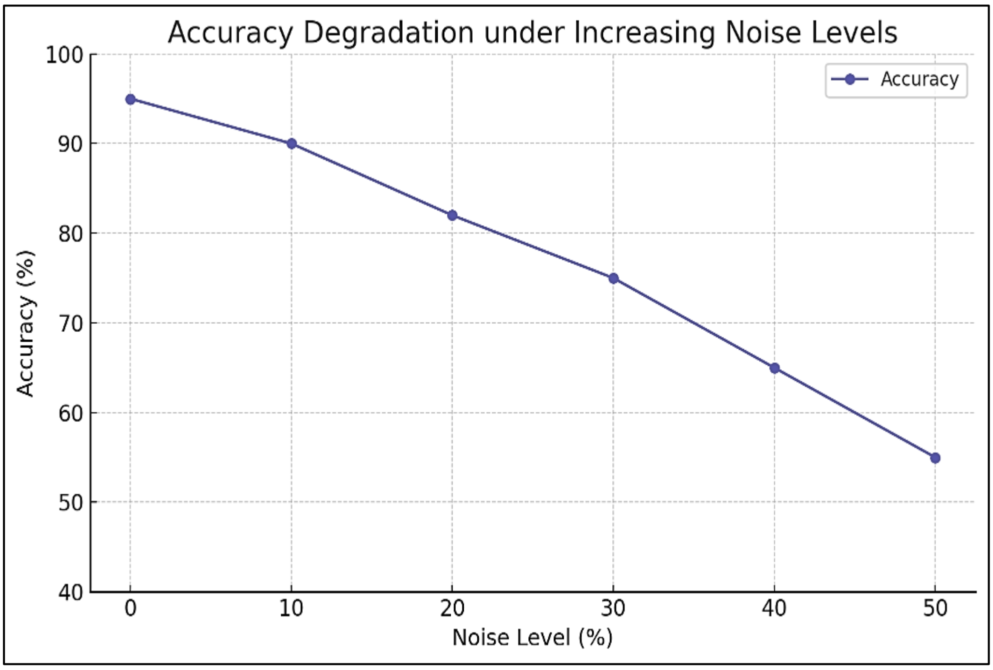

Figure4: Accuracy Degradation Under Increasing Noise Levels

5. Conclusion

In this study, a real-time emotion recognition HRI system is proposed, which utilizes deep learning models and brings emotional intelligence to robotic systems. Real-time high precision human emotion detection with the combination of multimodal information such as facial and speech expressions was exploited through this work, enabling a robot to sense human emotions more accurately and rapidly while in contact with human in real-world scenarios.

The results of experiments demonstrate that deep learning-based methods, especially CNNs for visual data and LSTM-based architectures for sequential audio data, can obtain high accuracy with a small delay, so they are suitable for real-time HRI. The integration of such models into robotic systems facilitates more natural, adaptive, and captivating HRI and contributes, in the long term, to the development of truly social and empathetic agents.

Although the results are promising, it is difficult to generalize to various subjects, to be executed in real environmental noise and unconstrained environment, and to eliminate/reduce bias in dataset. Future work should be investigated to improve the robustness by further cross-cultural emotion databases, and by more advanced fusion strategies for on-line multimodal information processing.In the long term this work represents a significant stride towards enabling emotional intelligence in robotic platforms, potentially allowing for their widespread deployment in venues such as healthcare, education, and personal assistance, where empathy and user adaptation are crucial.

Funding

None

ACKNOWLEDGEMENT

None

Declaration of Using AI Tools

The authors confirm that artificial intelligence (AI)–assisted tools were used solely to enhance the linguistic quality of the manuscript, including grammar correction and language clarity.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

References

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, pp. 98–125, 2017.

[2] P. Barros, D. Jirak, and S. Wermter, “Multimodal emotional state recognition using facial, vocal, and gestural cues in social robots,” *IEEE Trans. Cogn. Dev. Syst.*, vol. 13, no. 2, pp. 370–382, 2021.

[3] A. Dzedzickis, M. Tamosiunaite, and A. Gudi, “Human emotion recognition: Review of sensors and methods,” *Sensors*, vol. 20, no. 3, p. 592, 2020.

[4] A. Bera, S. Kim, T. Randhavane, A. Pratapa, and D. Manocha, “The socially invisible robot: Navigation in the social world using robot’s emotional intelligence,” *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 324–331, 2019.

[5] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2020.

[6] B. P. Majumder, S. Li, J. Ni, and J. McAuley, “Generating personalized recipes from historical user preferences,” in *Proc. Conf. Empirical Methods in Natural Language Processing and Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5976–5982.

[7] F. Noroozi *et al.*, “Survey on emotional body gesture recognition,” *IEEE Trans. Affect. Comput.*, vol. 9, no. 3, pp. 325–345, 2018.

[8] X. Chen, L. Zhao, and H. Liu, “Real-time emotion recognition system based on lightweight CNN for social robots,” *Sensors*, vol. 22, no. 14, p. 5302, 2022. [Online]. Available: https://doi.org/10.3390/s22145302

[9] D. Kollias *et al.*, “Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond,” *Int. J. Comput. Vis.*, vol. 127, no. 6, pp. 620–641, 2020.

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. Schuller, “Deep architecture enhancement for cross-corpus speech emotion recognition,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 1–13, 2020.