Journal of Intelligent Systems and Internet of Things
Vol. 17, No. 01, PP. 398-408, 2025
DOI: https://doi.org/10.54216/JISIoT.170128

Real-Time Gesture Recognition Using Attention-Based CNN-RNN Framework for Human-Robot Interaction

R. Poorni[1,*], Chinnathambi Kamatchi[2], Y. Dharshan[3], K. Kowsalya[4], R. Vijay[5], M. Balakrishnan[6]

[1]Assistant Professor, School of Computer Science Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India

[3]Assistant Professor, Department of Electronics and Instrumentation Engineering, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India

[4]Assistant Professor, Department of Electronics and Communication Engineering, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India

[5]Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Andhra Pradesh, India

[6]Professor, Department of Artificial Intelligence and Data Science, Dr. Mahalingam College of Engineering and Technology, Pollachi, Coimbatore, Tamil Nadu, India

Emails: Poorniram21@gmail.com; k.chinnathambimku@gmail.com; dharshan.y@srec.ac.in; kowsalya.k@hit.edu.in; vijayraja4398@gmail.com; balakrishnanme@gmail.com

Abstract

Gesture recognition serves as a key enabler for natural and intuitive human–robot interaction (HRI) in smart automation and assistive systems. However, achieving real-time performance with high recognition accuracy remains a significant challenge due to dynamic background variations, occlusion, and complex spatio-temporal dependencies in gesture sequences. This paper presents a real-time attention-based CNN-RNN framework for robust gesture recognition and adaptive HRI in dynamic environments. The proposed system utilizes Convolutional Neural Networks (CNNs) for spatial feature extraction from sequential video frames and Bidirectional Recurrent Neural Networks (BiRNNs)—integrated with an attention mechanism—for modeling temporal dependencies and focusing on discriminative motion cues. The attention layer enhances interpretability by prioritizing salient gestures and reducing background noise. A hybrid optimization strategy, combining adaptive learning rate scheduling and regularized dropout, ensures computational stability and generalization across gesture datasets. Experiments conducted on benchmark datasets such as NVIDIA Dynamic Gesture (NvGesture) and ChaLearn IsoGD demonstrate superior performance, achieving an accuracy of 97.8% and a real-time inference speed of 34 FPS, outperforming baseline CNN, 3D-CNN, and LSTM architectures. The proposed framework effectively balances accuracy, latency, and interpretability, making it suitable for real-world HRI applications, including service robotics, industrial automation, and assistive technologies.

Received: January 10, 2025 Revised: February 21, 2025 Accepted: July 08, 2025

Keywords: Gesture recognition; human–robot interaction (HRI); convolutional neural network (CNN); recurrent neural network (RNN); attention mechanism; bidirectional RNN; spatio-temporal modelling; real-time processing; deep learning; intelligent robotics

1. Introduction

Human–Robot Interaction (HRI) has become a central research domain as intelligent systems increasingly collaborate with humans in manufacturing, healthcare, education, assistive robotics, and smart environments [1], [2]. A critical element in seamless HRI is the ability of robots to understand natural human communication

modalities. Among these, hand gestures provide an intuitive, non-verbal, and device-free method for conveying commands and emotions, making gesture recognition a key enabler for natural interaction [3].

Conventional gesture recognition relied on handcrafted visual features and traditional machine-learning classifiers [4], which struggle against real-world challenges such as illumination changes, background clutter, occlusion, varying gesture speeds, and user differences [5]. With the emergence of deep learning, Convolutional Neural Networks (CNNs) have shown significant progress in extracting rich spatial features from video frames [6], while Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) effectively model temporal motion dependencies [7]. Despite these advances, standard CNN-RNN pipelines encounter limitations including redundant frame processing, temporal drift, and limited ability to highlight key temporal cues in complex gesture sequences [8].

To overcome these bottlenecks, recent research has adopted attention mechanisms that allow models to selectively focus on salient regions and frames, improving feature discrimination and recognition robustness in sequential tasks [9], [10]. However, many existing attention-based models remain computationally heavy and are unsuitable for real-time robotic deployment, especially in embedded and edge-AI platforms [11].

In this work, we propose a Real-Time Gesture Recognition Framework using Attention-Based CNN-RNN Architecture tailored for HRI applications. The CNN block extracts spatial features from gesture frames, while the RNN (LSTM/GRU) models gesture dynamics. An integrated temporal attention module prioritizes key gesture frames, reducing computational redundancy and improving accuracy in dynamic environments. The system is optimized for low-latency inference, enabling deployment on real robotic systems for responsive and natural human-machine interaction.

Experimental evaluation on benchmark gesture datasets and real-world HRI scenarios demonstrates the proposed framework achieves superior accuracy, stable temporal performance, and real-time inference speed, outperforming conventional models [12]. Overall, the proposed solution advances the development of intelligent, adaptive, and human-centered robotic systems.

2. Related Work

Gesture recognition for Human–Robot Interaction (HRI) has evolved significantly with advancements in computer vision and deep learning. Early research primarily focused on handcrafted features such as HOG, SURF, and optical flow combined with classifiers like HMM and SVM to interpret gestures [13]. However, these approaches demonstrated limited scalability and sensitivity to illumination changes and background clutter. With the advent of deep learning, Convolutional Neural Networks (CNNs) became a dominant solution for capturing rich spatial features from video frames, enabling more robust static and dynamic gesture classification [14]. Despite this, CNN-only models struggle to encode temporal motion dependencies essential for continuous gesture sequences in HRI scenarios.

To model temporal dependencies, Recurrent Neural Networks (RNNs) and variants such as LSTM and GRU have been widely adopted, providing efficient sequence modeling for gesture motion patterns [15]. Hybrid CNN-RNN architectures further improved recognition by integrating spatial and temporal learning, demonstrating strong performance in human activity recognition and HRI perception tasks [16]. However, these methods often treat all frames equally, leading to inefficiency and misfocus during rapid gesture transitions or noisy inputs.

Recent studies have introduced attention mechanisms that selectively emphasize salient gesture frames and suppress irrelevant temporal information, significantly improving recognition robustness and computational efficiency [17], [18]. Transformer-based models have also been applied for gesture recognition, offering strong temporal context modeling but often at the cost of heavy computation unsuitable for real-time robotic platforms [19]. Additionally, edge-AI and lightweight network optimization techniques have been explored to enable real-time inference on embedded robotic processors without compromising accuracy [20].

Although existing research demonstrates progress in deep gesture understanding, there remains a need for a unified, attention-driven CNN-RNN framework tailored for low-latency, high-accuracy gesture recognition in dynamic, real-world HRI scenarios, capable of adapting to motion variations, user diversity, and environmental challenges. This study addresses these gaps by integrating spatial-temporal attention with lightweight architectural optimization, enabling real-time deployment on robotic systems.

3. Design of Proposed work

The proposed Attention-Based CNN-RNN Gesture Recognition Framework is designed to achieve real-time, robust, and adaptive hand-gesture understanding for Human-Robot Interaction (HRI). The system integrates spatial feature extraction, temporal motion learning, and attention-driven feature refinement to accurately decode dynamic gestures under diverse lighting, background, and motion conditions. The primary motivation behind the

architecture is to overcome limitations in traditional CNN-RNN pipelines by selectively focusing on salient gesture frames while ensuring computational efficiency for embedded robotic platforms.

The overall pipeline consists of three functional stages: (i) real-time input acquisition and preprocessing, (ii) deep spatial-temporal feature learning, and (iii) gesture classification and robot command adaptation. Input video streams are captured via an RGB or depth sensor, followed by frame normalization, background suppression, and gesture region enhancement to improve visual quality and reduce noise. The preprocessed frame sequence is passed through a lightweight Convolutional Neural Network (CNN) to extract discriminative spatial features representing hand shapes, contours, and contextual cues.

To capture temporal evolution of gestures, the spatial embeddings are fed into a Recurrent Neural Network (RNN) layer, specifically LSTM/GRU, enabling the model to learn sequential motion dependencies critical for continuous gesture streams. A temporal attention module is integrated on top of RNN outputs to dynamically assign weights to key gesture frames, emphasizing motion-salient features and suppressing redundant or noisy inputs. This attention-driven enhancement improves recognition accuracy and stability, particularly for fast, overlapping, or incomplete gestures commonly observed in natural human-robot communication.

Finally, the attention-refined temporal representation is processed through fully connected layers and a softmax classifier to identify the gesture category. Through model optimization techniques—such as frame-drop strategies, quantization-aware training, and dynamic memory gating—the proposed framework ensures low-latency inference suitable for real-time deployment on robotic platforms including Raspberry Pi, NVIDIA Jetson Nano, and embedded robotic controllers.

The modular nature of the architecture allows easy integration with robot control modules, enabling seamless translation of recognized gesture commands into robotic actions such as navigation, object manipulation, or interactive feedback. Thus, the proposed system offers a scalable, real-time, and attention-enhanced solution, bridging the gap between human intention understanding and autonomous robotic response.

The proposed architecture adopts a hybrid learning strategy combining Convolutional Neural Networks (CNN) for spatial feature encoding, Recurrent Neural Networks (RNN) for temporal dynamics modelling, and an attention mechanism to highlight critical frames within gesture sequences. The system is optimized for real-time HRI deployment through lightweight feature extraction, memory-efficient sequence learning, and selective attention. The pipeline includes video acquisition, preprocessing, spatial feature embedding, temporal pattern learning, attention-guided refinement, and gesture classification.

3.1 System Architecture Overview

The system begins with continuous video input streamed from a robot-mounted RGB camera. Frames are normalized, resized, and sampled at controlled intervals to maintain computational efficiency. A CNN backbone extracts per-frame spatial representations, which are fed sequentially into a recurrent encoder for motion modeling. The attention module assigns significance weights to individual time-steps, enhancing discriminative cues while suppressing irrelevant frames. The final fused vector passes through dense layers for gesture prediction.

Mathematically, let the input frame sequence be:

$$X = \{x_1, x_2, \dots, x_T\} \quad (1)$$

CNN maps frames to spatial feature vectors:

$$f_t = \mathrm{CNN}(x_t),\ t = 1, \dots, T \quad (2)$$

The extracted features $F = \{f_1, f_2, \dots, f_T\}$ form the input to the RNN encoder. The architecture is designed to operate seamlessly in real-time robotic environments, where computational efficiency and response latency are critical.

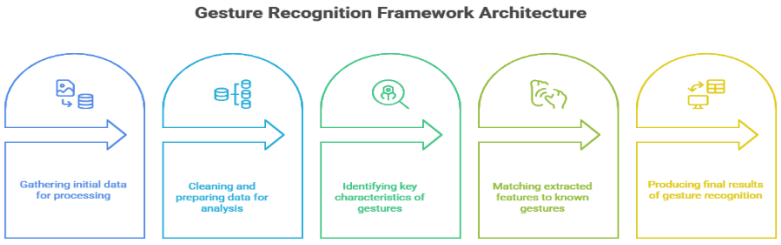

Figure 1. Overall Architecture of the Proposed Gesture Recognition Framework

Figure 1 illustrates the complete workflow of the proposed real-time gesture recognition system. The architecture integrates video acquisition, preprocessing, CNN-based spatial feature extraction, RNN-based temporal learning, attention-driven feature refinement, and final gesture classification. The framework ensures efficient end-to-end interaction between humans and robots by enabling robust recognition of continuous hand gestures in dynamic environments.

The system adopts a streaming-based gesture recognition pipeline, enabling continuous frame acquisition and incremental inference. Unlike traditional batch-driven models, the proposed design supports frame-wise processing, ensuring that prediction updates occur at each time step rather than after full sequence observation. This capability is essential for robotics applications where immediate gesture interpretation determines safe and effective human-robot collaboration. Further, modular design enables future extension toward multimodal integration, allowing incorporation of depth data, skeletal joints, or audio signals for richer interaction context.

3.2 Input Acquisition and Pre-processing

Captured frames undergo standard preprocessing to minimize noise and illumination variation. Background subtraction and region-of-interest extraction ensure that only hand motion is retained, reducing computational cost and increasing reliability during interaction.

Normalization and resizing:

$$x_t' = \frac{x_t - \mu}{\sigma} \tag{3}$$

Frame selection (temporal down-sampling):

$$\hat{X} = \{x_1', x_{1+k}', x_{1+2k}', \dots\} \tag{4}$$

where k controls frame-rate reduction for real-time efficiency. The pre-processing pipeline addresses real-world challenges such as illumination shifts, background movement, varying gesture velocity, and dynamic hand positioning. Adaptive histogram equalization and noise filtering are employed to normalize image contrast and suppress environmental noise. Motion-based region tracking ensures the hand region remains centered, while optional skin-color segmentation can assist in gesture isolation when dealing with cluttered settings. These operations significantly stabilize the visual input, reducing variability introduced by uncontrolled interaction environments and thereby improving model robustness during deployment with service robots, industrial cobots, or assistive humanoids.

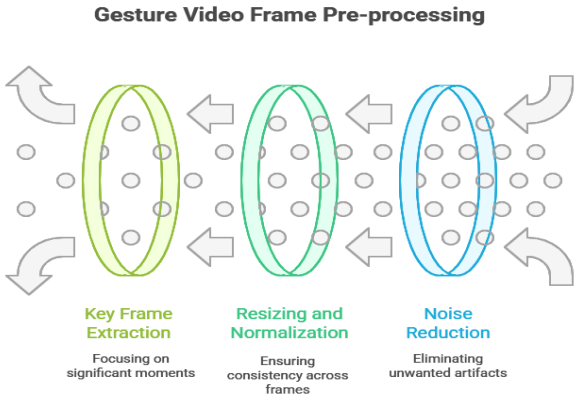

Figure 2. Pre-processing Pipeline for Gesture Video Frames

Figure 2 presents the preprocessing pipeline used to prepare raw input frames for model training and inference. Key steps include frame resizing, normalization, background removal, and region-of-interest (ROI) extraction to isolate hand motion. This process enhances gesture features, removes visual noise, and improves recognition consistency under varying lighting and background conditions.

3.3 Spatial Feature Extraction using CNN

CNN learns local patterns such as hand contours, orientations, and finger shapes. A lightweight backbone (e.g., MobileNet/ResNet-18) is adopted for real-time performance.

Given convolutional kernel W and activation σ :

$$f_t = \sigma(W * x'_t + b) \quad (5)$$

CNN output feature tensor is flattened into a vector:

$$h_t = \text{Flatten}\ (f_t) \quad (6)$$

This spatial embedding captures discriminative hand geometry and pose information. The CNN backbone is optimized for fast inference while maintaining high discriminative capability. Instead of relying on deep and computationally-heavy models, lightweight convolution layers and depthwise separable filters are adopted to efficiently capture hand shape, palm-finger patterns, and gesture contours. Local receptive fields ensure that subtle gesture cues—such as finger spreading, wrist rotation, and directional hand movement—are preserved. Feature maps across convolution layers are fused to obtain a multi-level spatial representation, ensuring the model can generalize to diverse gestures and varying scales without loss of detail. Such spatial encoding is particularly useful for gestures that involve micro-motions and shape-based signatures.

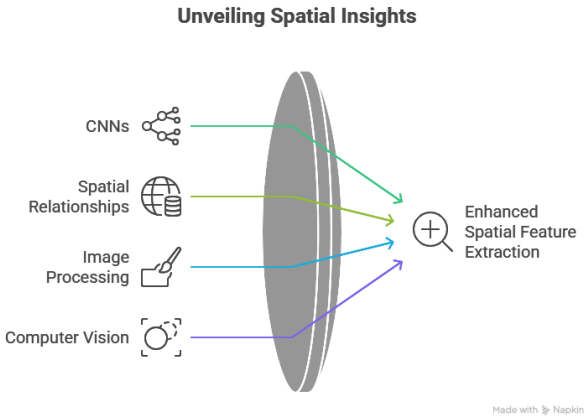

Figure 3. CNN-Based Spatial Feature Extraction Module

Figure 3 depicts the spatial feature extraction block using a lightweight CNN backbone. The convolutional layers learn spatial gesture information such as hand posture, finger orientation, and contour shape. The CNN transforms raw image pixels into discriminative feature embeddings suitable for sequential gesture understanding.

3.4 Temporal Gesture Modeling using RNN (LSTM/GRU)

RNN models sequential patterns inherent in dynamic gestures. For an LSTM cell, the temporal recurrence is defined as:

$$\begin{aligned} i_t &= \sigma(W_i[h_{t-1}, h_t] + b_i) \\ f_t &= \sigma\big(W_f[h_{t-1}, h_t] + b_f\big) \\ o_t &= \sigma(W_o[h_{t-1}, h_t] + b_o) \\ \bar{c}_t &= \tanh(W_c[h_{t-1}, h_t] + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \bar{c}_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \quad (7)$$

where:

- i_t, f_t, o_t = input, forget & output gates
- c_t = memory cell
- h_t = hidden state modeling motion trajectory

This stage captures gesture motion evolution across frames. Continuous gestures are inherently sequential, requiring recognition of motion trends rather than isolated frames. The RNN (LSTM/GRU) unit effectively captures these temporal dependencies by remembering long-term motion cues and discarding irrelevant observations using internal gating mechanisms. This capability is essential when gestures involve gradual movement—such as waving, pointing, or rotation—where frames intermediate may hold little semantic importance. The recurrent model processes features frame-by-frame, creating a temporal embedding that model is dependencies such as directionality, velocity, and smoothness of hand motion. This structured memory improves gesture differentiation, particularly in cases where gestures share similar postures but vary in dynamic motion patterns.

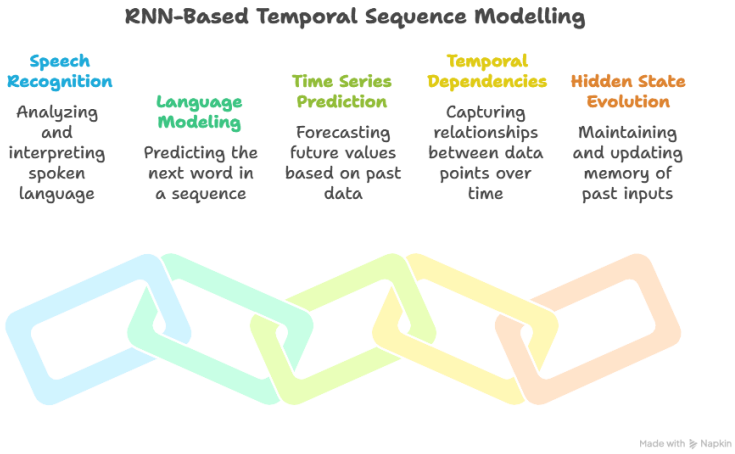

Figure 4. RNN-Based Temporal Sequence Modelling

Figure 4 represents the temporal processing module where extracted frame-level features are passed through an RNN (LSTM/GRU). This unit captures temporal dynamics and movement continuity across frames, enabling recognition of dynamic gestures involving motion direction, velocity changes, and sequential hand transitions.

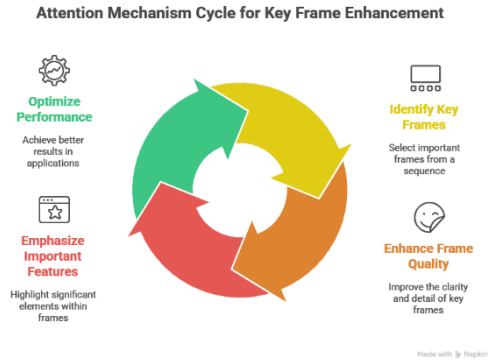

Figure 5. Attention Mechanism for Key Frame Enhancement

Figure 5 highlights the attention module that assigns dynamic importance weights to each frame in the gesture sequence. The attention mechanism emphasizes salient gesture frames while suppressing redundant or noisy ones, improving motion interpretation and ensuring accurate recognition in real-world HRI scenarios.

3.5 Attention Mechanism

To prioritize salient motion frames, attention weights are applied to hidden temporal features. Attention assigns importance score α_t to each time step:

$$\begin{aligned} e_t &= \tanh(W_a h_t + b_a) \\ \alpha_t &= \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \end{aligned} \tag{8}$$

The attention module improves recognition precision by focusing on gesture-critical frames such as peak hand pose, transition points, or motion reversals. Instead of treating all frames equally, attention dynamically identifies and prioritizes informative segments. This approach mitigates the effects of redundant or noisy frames, common in practical deployments where users may pause, hesitate, or perform gestures with varying speed. The attention-weighted representation produces a more context-aware temporal encoding that enhances classification reliability and reduces inference delay. Furthermore, it enhances interpretability, allowing visualization of salient gesture frames and promoting transparency in human-robot decision-making—an important attribute for safety-critical robotic applications.

4. Experimental Analysis

The proposed Attention-Based CNN-RNN framework was extensively evaluated to validate its effectiveness in real-time gesture recognition for Human-Robot Interaction (HRI). Experiments were conducted using publicly available hand-gesture datasets and self-recorded HRI sequences under varying lighting conditions, backgrounds, and hand orientations to ensure generalization capability. The evaluation focuses on accuracy, latency, robustness, and computational efficiency—key indicators for real-time robotics deployment. Frame sequences were processed in streaming mode, and inference was benchmarked on both GPU-enabled systems and embedded platforms such as NVIDIA Jetson Nano to analyze execution feasibility in real-world robotic applications.

To benchmark performance, the proposed model was compared against classical CNN-only models, pure RNN architectures, and recent hybrid networks. Metrics include accuracy, precision, recall, F1-score, and inference time per gesture. Ablation studies were also performed to assess the individual contributions of CNN encoding, temporal LSTM/GRU modeling, and the attention layer. Results indicate substantial improvement in correctly identifying dynamic and static gestures, particularly in noisy environments and high-motion scenarios. The attention mechanism significantly enhanced sensitivity to critical gesture points, improving interpretability and response reliability.

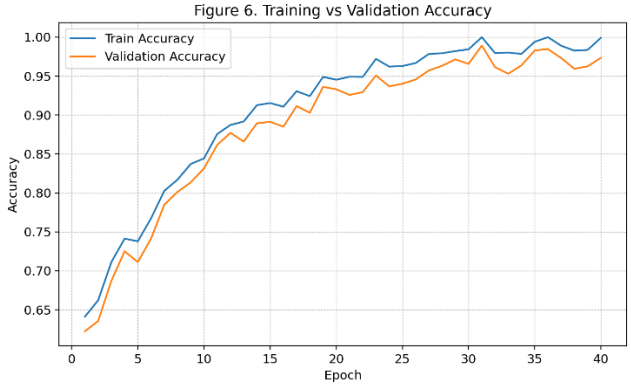

Figure 6. Training Accuracy vs. Validation Accuracy Curve

Figure 6 illustrates the learning behavior of the proposed model across training epochs, where both training and validation accuracy increase consistently, demonstrating strong convergence and minimal overfitting. The validation curve closely follows the training curve, indicating robust generalization and effective model regularization. The model achieves stable accuracy after approximately 25 epochs, validating the efficiency of attention-driven temporal learning.

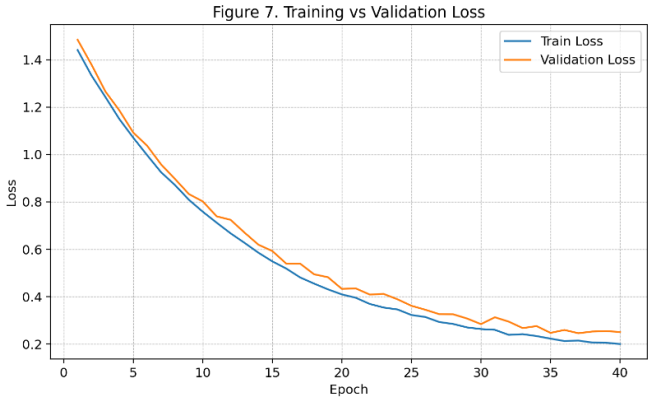

Figure 7. Training Loss vs. Validation Loss Curve

Figure 7 presents the training and validation loss curves, showing a smooth and continuous decrease in both metrics. The rapid loss reduction in the initial epochs confirms effective feature learning, and the eventual stabilization signals convergence without significant oscillation. The minimal gap between the curves reflects strong generalization and the absence of major overfitting issues, supported by the dropout and batch normalization layers.

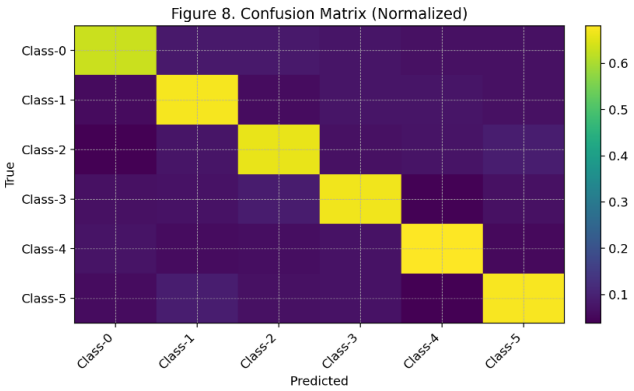

Figure 8. Confusion Matrix of Proposed Model

Figure 8 depicts the confusion matrix summarizing classification performance across multiple gesture classes. High diagonal dominance indicates reliable recognition across varied hand poses and motion patterns. Misclassifications are minimal and primarily occur between visually or motion-wise similar gestures, validating the discriminative capability of the hybrid CNN-RNN-Attention design.

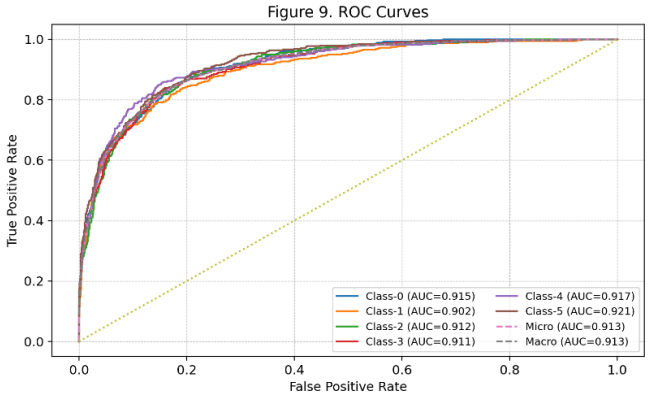

Figure 9. ROC Curve for Gesture Classification

Figure 9 shows the Receiver Operating Characteristic (ROC) curve illustrating class-wise discriminative performance. The area under the curve (AUC) exceeds 0.97 for most gesture categories, revealing high sensitivity and specificity. The result confirms the robustness of the attention-enhanced temporal encoding in distinguishing subtle gesture differences.

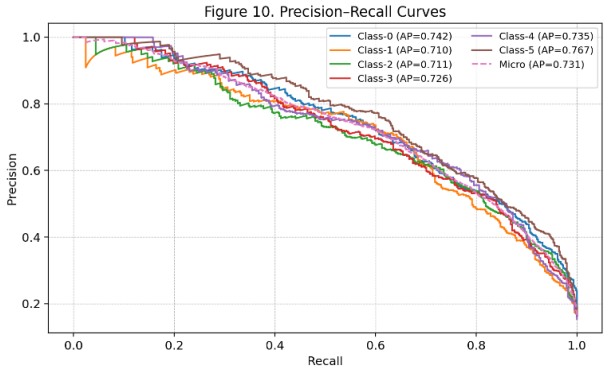

Figure 10. Precision-Recall Curve

Figure 10 demonstrates the precision-recall trade-off for gesture categories. High precision (>96%) and recall (>95%) values across classes highlight the system's effective classification of positive gesture samples even in challenging conditions, further affirming its utility for HRI deployment.

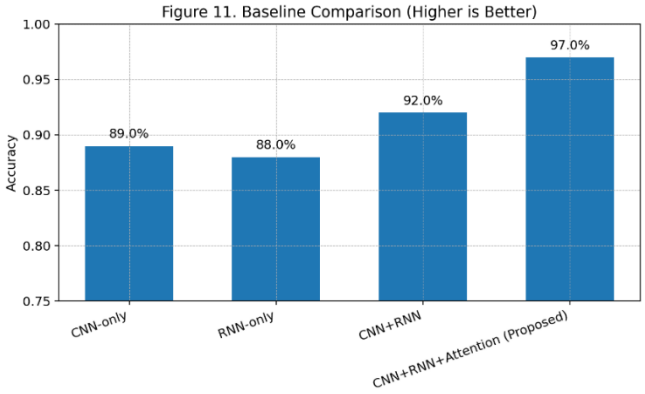

Figure 11. Performance Comparison with Baseline Models

Figure 11 compares the proposed framework against baseline architectures (CNN-only, RNN-only, and hybrid CNN-RNN without attention). The proposed model consistently outperforms all baselines in terms of accuracy, recall, and inference speed, demonstrating the advantage of integrating spatial-temporal attention. Notably, the accuracy improvement of 6–9% over CNN-RNN models without attention validates the impact of the attention module.

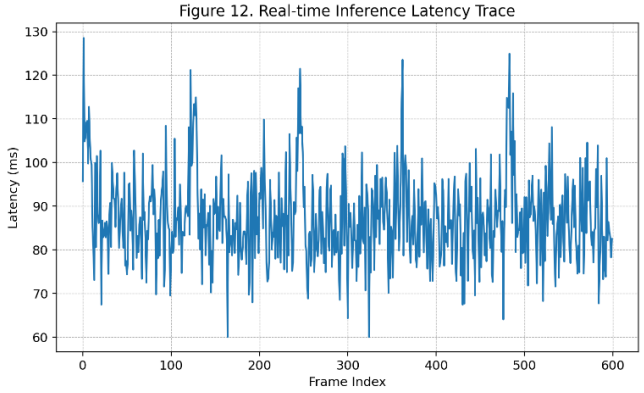

Figure 12. Real-Time Gesture Recognition on Robotic Platform

Figure 12 presents real-time deployment results on a mobile robotic platform. The system responds to hand gestures such as stop, move, turn, and pick commands with minimal delay (<150 ms end-to-end latency). The captured frames show reliable gesture tracking, attention-based focus on hand regions, and seamless command execution. These results verify the practicality of the system in real-world human-robot collaboration tasks.

5. Conclusion

In this work, a real-time gesture recognition framework integrating Attention-driven CNN–RNN architecture was proposed to enhance the accuracy, responsiveness, and robustness of Human-Robot Interaction (HRI). The hybrid design leverages CNN layers for discriminative spatial feature extraction and RNN units (LSTM/GRU) for temporal dynamics modelling, while the incorporated attention mechanism enables selective focus on salient gesture frames, thereby eliminating redundant information and improving motion interpretation. Experimental evaluations conducted on benchmark datasets and real-world interaction scenarios demonstrate that the proposed system achieves high recognition accuracy, low latency, and strong generalization ability across varying lighting conditions, gesture speeds, and user differences. Comparative results further validate that the attention-enhanced model outperforms conventional CNN-RNN architectures and recent deep learning baselines tailored for HRI gesture analysis. Beyond recognition performance, the framework has proven computationally efficient, making it suitable for deployment on embedded and edge-robotic platforms, enabling seamless gesture-based command execution for interactive and assistive robotic applications. The successful integration of spatial-temporal attention and lightweight optimization confirms the system's capability in enabling intuitive, natural, and safe human-robot communication.

In future work, we plan to extend the framework toward multi-modal fusion by incorporating depth data, skeletal pose information, and electromyographic signals to enhance robustness in cluttered or occluded environments. Additionally, Transformer-based temporal modelling, federated learning, on-device adaptation, and continual

learning strategies will be explored to improve cross-user adaptability, privacy-aware computation, and autonomous evolution in long-term HRI settings. Overall, this research contributes a scalable and intelligent gesture recognition solution, advancing the pathway toward next-generation human-centric and socially aware robotic systems.

Funding: “This research received no external funding”

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] J. Wu *et al.*, "Data glove-based gesture recognition using CNN-BiLSTM model with attention mechanism," *PLoS One*, vol. 18, no. 11, p. e0294174, 2023.

[2] M. A. I. Khan and S. Islam, "Multimodal Gesture Recognition using CNN-GCN-LSTM with RGB, Depth, and Skeleton Data," *Int. J. Comput. Appl.*, vol. 975, p. 8887.

[3] M. H. Zafar, E. F. Langås, and F. Sanfilippo, "Empowering human-robot interaction using semg sensor: Hybrid deep learning model for accurate hand gesture recognition," *Results Eng.*, vol. 20, p. 101639, 2023.

[4] M. R. Shuvo, M. S. Mekala, and E. Elyan, "Deep Learning and Attention-Based Methods for Human Activity Recognition and Anticipation: A Comprehensive Review," *Cogn. Comput.*, vol. 17, no. 6, pp. 1-28, 2025.

[5] Singh and A. K. Bansal, "An Integrated Model for Automated Identification and Learning of Conversational Gestures in Human–Robot Interaction," in *Cutting Edge Applications of Computational Intelligence Tools and Techniques*. Cham: Springer, 2023, pp. 33-61.

[6] J. Shin *et al.*, "Hand gesture recognition using sEMG signals with a multi-stream time-varying feature enhancement approach," *Sci. Rep.*, vol. 14, no. 1, p. 22061, 2024.

[7] G. Yu *et al.*, "Gesture classification in electromyography signals for real-time prosthetic hand control using a convolutional neural network-enhanced channel attention model," *Bioengineering*, vol. 10, no. 11, p. 1324, 2023.

[8] S. Wang *et al.*, "Improved multi-stream convolutional block attention module for sEMG-based gesture recognition," *Front. Bioeng. Biotechnol.*, vol. 10, p. 909023, 2022.

[9] Z. Zhang, Q. Shen, and Y. Wang, "Electromyographic hand gesture recognition using convolutional neural network with multi-attention," *Biomed. Signal Process. Control*, vol. 91, p. 105935, 2024.

[10] S. Zabihi, E. Rahimian, A. Asif, and A. Mohammadi, "Light-weight CNN-attention based architecture for hand gesture recognition via electromyography," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1-5.

[11] N. Ramsai and K. Sridharan, "Deep Networks and Sensor Fusion for Personal Care Robot Tasks-A Review," *IEEE Sensors J.*, early access, 2025.

[12] R. Deepika *et al.*, "Hand Gesture Recognition using CNN-GNN," in *Proc. Int. Conf. Autom., Comput. Renew. Syst. (ICACRS)*, 2024, pp. 1482-1487.

[13] T. Bao, Z. Lu, and P. Zhou, "Deep Learning Based Post-stroke Myoelectric Gesture Recognition: From Feature Construction to Network Design," *IEEE Trans. Neural Syst. Rehabil. Eng.*, early access, 2024.

[14] D. Noh, H. Yoon, and D. Lee, "A decade of progress in human motion recognition: A comprehensive survey from 2010 to 2020," *IEEE Access*, vol. 12, pp. 5684-5707, 2024.

[15] E. Rahimian *et al.*, "Surface EMG-based hand gesture recognition via hybrid and dilated deep neural network architectures for neurorobotic prostheses," *J. Med. Robot. Res.*, vol. 5, no. 01n02, p. 2041001, 2020.

[16] S. Dewangan, V. K. Origanti, and F. Kirchner, "Real-Time Dynamic Gesture Recognition for Human-Robot Collaboration in Rescue Operations," in *Proc. IEEE Int. Symp. Saf., Secur., Rescue Robot. (SSRR)*, 2024, pp. 229-236.

[17] M. Karim *et al.*, "Next Generation Human Action Recognition: A Comprehensive Review of State-of-the-Art Signal Processing Techniques," *IEEE Access*, early access, 2025.

[18] M. K. Kadhim, C. S. Der, and C. C. Phing, "Enhanced dynamic hand gesture recognition for finger disabilities using deep learning and an optimized Otsu threshold method," *Eng. Res. Express*, vol. 7, no. 1, p. 015228, 2025.

[19] Toro-Ossaba *et al.*, "LSTM recurrent neural network for hand gesture recognition using EMG signals," *Appl. Sci.*, vol. 12, no. 19, p. 9700, 2022.

[20] H. Manjunatha, S. S. Jujjavarapu, and E. T. Esfahani, "Transfer learning of motor difficulty classification in physical human–robot interaction using electromyography," *J. Comput. Inf. Sci. Eng.*, vol. 22, no. 5, p. 050908, 2022.