International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

INDIAN SIGN LANGUAGE RECOGNITION AND COGNITIVE OPTIMIZATION USING A MULTIMODAL ROBOTIC SYSTEM

Shreeraj Desai[1*], Ghanshyam Kanojiya[2], Shobit Halse[3], Yeshwant Naik[4], Kanojia Mahendra[5] and Kunal Joshi[6]
[1,2,3,4,5,6]Department of Computer Science, Sheth. L.U.J. and Sir M.V. College, India
[1]shreerajd300@gmail.com, [2]kanojiyaghanshyam92@gmail.com, [3]shobithalse.research@gmail.com, [4]yeshwantnaik076@gmail.com@gmail.com, [5]kgkmahendra@gmail.com, [6]kunaljoshi007m@gmail.com
Corresponding author: Shreeraj N Desai, Department of Computer Science, Sheth. L.U.J. and Sir M.V. College, India, shreerajd300@gmail.com

ABSTRACT
Modern Intelligent Personal Assistants rely predominantly on voice-based interaction, creating systemic accessibility barriers for the Deaf and Hard-of-Hearing community while failing to provide active cognitive support for the general population. This study presents a proposed multimodal Socially Assistive Robot named "Zeus," designed to bridge these accessibility and wellness gaps through a distributed embodied Artificial Intelligence system. The research aims to engineer a cost-effective, real-time platform capable of interpreting Indian Sign Language for inclusive communication while simultaneously optimizing user cognitive states through empathetic interaction. The methodology employs a distributed client-server architecture to offload heavy computation from a Raspberry Pi 4 to an external server. The vision pipeline integrates MediaPipe landmarks with a Long Short-Term Memory network for dynamic gesture classification, while the wellness module utilizes a Random Forest classifier trained on EEG data to trigger personalized 10 Hz Alpha binaural beats. Experimental results demonstrate that the system achieves an ISL recognition accuracy of 99.85% within the evaluated dataset and a wellness classification accuracy of 91.7%, while maintaining an average system latency of 180 ms. These findings suggest that integrating affective computing with computer vision can successfully transform digital assistants into inclusive, wellness-oriented companions for both the Deaf and Hard-of-Hearing community and the general public.

Keywords: *Socially Assistive Robotics, Indian Sign Language Recognition, Deep Learning, Binaural Beats, Long Short-Term Memory, Distributed Edge Computing.*

1. INTRODUCTION
The rapid growth of Artificial Intelligence (AI) has completely changed how we interact with computers by moving us from typing on keyboards to speaking to assistants like Amazon Alexa and Apple Siri. These tools have become a normal part of daily life, but their design has created a major barrier for many people. According to the (World Health Organization, 2021), over 430 million people globally live with disabling hearing loss, and that number could reach 700 million by 2050. In India alone, the situation is critical, as estimates suggest that around 63 million people face significant hearing challenges. Despite these huge numbers, most smart assistants still rely almost entirely on voice, which creates a serious accessibility gap. Research by (Sta. Maria & Deja, 2024) shows that Deaf and Hard-of-Hearing (DHH) users face a Word Error Rate (WER) of up to 78% when trying to use these devices compared to just 18% for hearing users. This high failure rate means that modern technology is practically unusable for a large population that relies on Indian Sign Language (ISL) to communicate.

However, the limitations of current technologies extend beyond accessibility. For the general public, existing assistants operate as passive tools devoid of presence or empathy. It is argued that these systems lack emotional intelligence because they cannot detect a user's cognitive state or offer companionship (Ma et al., 2023). In an era where stress and social isolation are rising, a disembodied voice from a speaker fails to provide the connection humans need. Research indicates that physical embodiment and active gaze significantly improve user engagement and reduce feelings of loneliness compared to static devices (Pan & Law, 2020). Furthermore, technology has the potential to actively optimize mental well-being for everyone. Specifically, it was found that auditory interventions like 16 Hz binaural beats can significantly modulate brain activity to improve behavioral performance (Al-Shargie et al., 2022). While multiple frequencies have shown efficacy, this study specifically focuses on 10 Hz Alpha waves target relaxation and stress reduction. Right now, there is no single affordable system that combines inclusive communication for the deaf with active companionship and stress regulation for the general user. Most existing solutions are either too expensive because they need powerful industrial computers, or they are too simple and cannot understand moving gestures (Katoch et al., 2022). As a result, there is a clear need for a low-cost, embodied system that can handle real-time sign language while acting as a supportive companion for any user.

ISSN 2394 - 7780

This study addresses these problems by introducing the proposed model of a multimodal Socially Assistive Robot (SAR) named "Zeus." The main goal is to build a system that can understand ISL in real-time and automatically generate personalized binaural beats to optimize the user's cognitive state. By combining computer vision models with a wellness algorithm and an animatronic gaze mechanism, this research shows that capable and empathetic robots can be built using affordable hardware like the Raspberry Pi. The system uses a distributed architecture to handle the heavy processing on a separate server, which keeps the robot fast and responsive.

2. THEORETICAL BACKGROUND

Between 2020 and 2025, assistive technology evolved from simple translation tools into more complex, interactive systems. Early research in 2020 focused on identifying structural constraints in existing solutions. A major challenge in Indian Sign Language (ISL) research was the scarcity of video data, which necessitated reliance on static images for initial validation (Wadhawan & Kumar, 2020). Concurrently, studies indicated that popular voice assistants presented usability challenges for users with speech impairments due to rigid timing protocols (Masina et al., 2020). Research also established that for a robot to maximize engagement, it requires a physical presence and active eye contact rather than functioning solely as an audio interface (Pan & Law, 2020). To enhance the adaptability of intelligent assistants beyond fixed training data, Retrieval-Augmented Generation (RAG) using Dense Passage Retrievers was introduced, allowing AI to query external data without requiring extensive retraining (Lewis et al., 2020).

By 2021 and 2022, the focus shifted to refining gesture recognition capabilities. An accuracy of 98.4% was achieved by utilizing Sign Language Graph Convolution Networks (SL-GCN) to track skeletal joints, demonstrating the efficacy of graph-based methods for complex movements (Jiang et al., 2021). While an even higher accuracy of 99.6% was reached using Support Vector Machines (SVM) and Convolutional Neural Networks (CNNs), those systems were optimized primarily for static postures using the Bag-of-Visual-Words technique, limiting their applicability to continuous sign language (Katoch et al., 2022). To facilitate deployment on compact devices, skeletal tracking was combined with Gated Recurrent Units (GRU). It was demonstrated that these lightweight recurrent models could maintain 95% accuracy even on edge hardware (Subramanian et al., 2022). These methods build upon the sequence modeling principles originally established by the Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997).

Parallel to vision-based research, studies began validating auditory beat stimulation as a therapeutic intervention. Before a robotic system can respond to user stress, it must reliably detect cognitive states from physiological data. It was demonstrated that Random Forest ensembles could accurately classify mental states using EEG data by modeling non-linear feature interactions (Bird et al., 2019; Breiman, 2001). To apply this therapeutically, it is necessary to map specific auditory frequencies to desired brain states. Table 1 outlines the standard frequency ranges utilized to target specific cognitive outcomes.

Table 1. EEG Frequency Bands and Their Associated Cognitive States

EEG Band	Frequency Range (Hz)	Associated Cognitive State
Delta	0.3 – 4	Deep relaxation / Subconscious
Theta	4–8	Memory and relaxation
Alpha	8–13	Calm attention
Beta	13–30	Concentration and alertness

Based on these categories, experiments confirmed that 10 Hz Alpha beats effectively reduce stress in students (Shalforoushan & Golmakani, 2022; Jurvanen, 2020). A subsequent analysis of 1,436 patients confirmed that these binaural beats are more effective than standard music for stabilizing blood pressure and anxiety (Xiong et al., 2025). Additionally, it was noted that when robots respond with emotional intelligence, users report a stronger sense of connection (Hong et al., 2021).

In the most recent phase from 2024 to 2025, researchers focused on speed and real-time performance. Algorithms like FingerNet (ResNet-16) and Connectionist Temporal Classification (CTC) were optimized to recognize continuous signs with minimal latency (Meng et al., 2024; Zuo et al., 2024; Karthik et al., 2024). However, hardware trade-offs remain a consideration. It was observed that when running on a Raspberry Pi using standard Contour Detection algorithms, accuracy decreased to 90% under variable lighting conditions (Trisha et al., 2025). Although previous studies have advanced specific technologies such as real-time ISL recognition and binaural beat generation, these functions typically exist as separate systems. Currently, there is no single system that integrates gesture translation, gaze-based interaction, and automated wellness support into one interface. This research addresses this limitation by developing the proposed model "Zeus," a distributed

ISSN 2394 - 7780

social robot. By combining LSTM-based sign language recognition with algorithmic binaural beat synthesis, this study aims to provide a dual-purpose solution that facilitates communication and supports cognitive wellness for the DHH community.

3. PROPOSED WORK

This study utilizes a quantitative experimental design based on a distributed systems approach to develop a multimodal SAR. To facilitate this, the research utilized a comprehensive data strategy divided into visual and auditory domains. For ISL recognition, two specialized datasets were employed. First, a static corpus of the ISL alphabet was utilized, comprising approximately 710 high-resolution images for each of the 26 English letter classes (Wadhawan & Kumar, 2020). This served as a controlled environment to validate the geometric accuracy of the feature engineering pipeline. Second, for the primary objective of dynamic recognition, a "Curated ISL Word Corpus" was assembled targeting a vocabulary of 69 distinct ISL word signs. Each sample in this corpus consists of a video clip with an approximate duration of 2 seconds recorded at 25 frames per second by multiple participants to ensure variance. To bridge the domain gap between clean training data and real-world variance, this dynamic corpus was expanded by a factor of five using a temporally-consistent augmentation strategy implemented via the Albumentations library (Buslaev et al., 2020). Simultaneously, the wellness module was developed using the core cognitive state classification model trained on the publicly available EEG Emotion Dataset (Mohsen, 2021), comprising 2,132 samples where each sample contained 2,548 pre-extracted features. To classify these high-dimensional features into cognitive states (Focused, Relaxed, Stressed) derived from established affective mappings in the circumplex model (Russell, 1980), the system utilizes a Random Forest (Breiman, 2001) ensemble classifier. This algorithm was selected over deep neural networks due to its superior performance on tabular EEG feature sets and robustness against overfitting on smaller sample sizes. Inclusion criteria required gestures to include upper-body skeletal motion, while frames with motion blur were excluded.

The system architecture follows a distributed client–server model, as shown in Figure 1. The Raspberry Pi 4 functions as the edge client, capturing visual and auditory inputs and handling physical actuation, while tasks that require higher processing power are performed on an external server. Data is streamed from the client to the server via TCP/IP sockets, and high-level control commands are returned to enable responsive interaction. To ensure reliability, the system includes a basic fallback protocol where the robot reverts to a local text-based interface if the network connection to the server is lost.

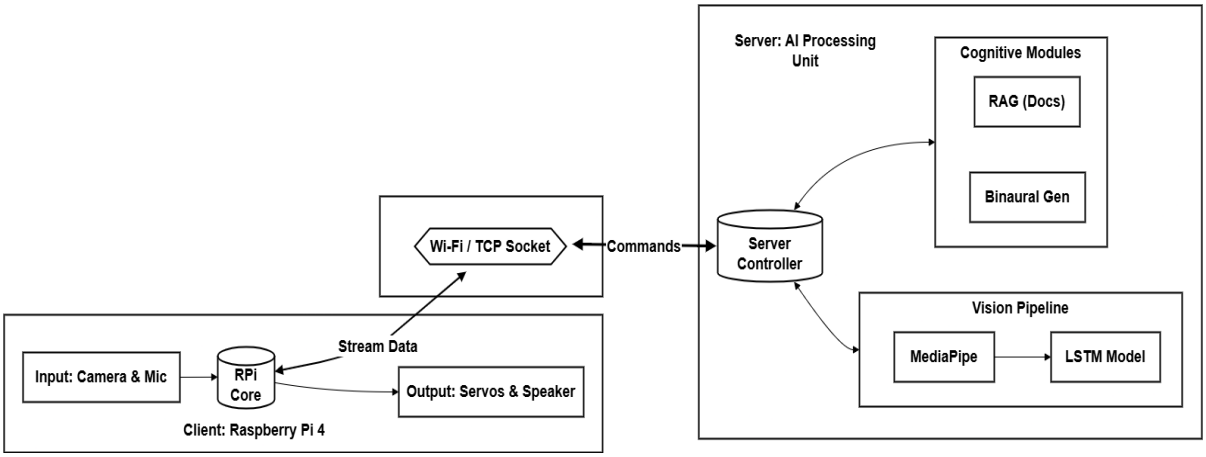

Figure 1: Distributed client-server architecture illustrating the data flow between the Raspberry Pi client and the external AI server.

The software pipeline integrates computer vision frameworks with distributed logic. In the preprocessing phase, raw video data streamed from the Raspberry Pi is normalized to a resolution of 640 by 480 pixels. MediaPipe (Lugaresi et al., 2019) extracts 21 hand landmarks. To reduce bias related to the user's distance from the camera, coordinates (x) are normalized using the Z-score formula (Kreyszig, 1979; Patro & Sahu, 2015), defined in Equation 1.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

In this equation, μ represents the mean and σ represents the standard deviation. Additionally, to convert categorical gesture labels into a machine-readable format, a label encoding function $f(l)$ is mapped to an integer vector, as shown in Equation 2. This technique assigns a unique integer to each class label (Potdar et al., 2017).

ISSN 2394 - 7780

$$y = f(l) \in \{0, 1, \cdots, K-1\} \quad (2)$$

Here, K represents the total number of gesture classes. This deterministic mapping ensures consistent target values during the supervised training phase. Following preprocessing, the system executes the operational logic detailed in Figure 2. The architecture functions as a parallel state machine that monitors two concurrent input streams: visual data from the webcam and auditory signals from the microphone. If the LSTM confidence score exceeds an 80% threshold or voice activity is detected, the system transitions from an 'Idle' state to an 'Active' state. In the Idle state, the matrix display renders a static line to conserve power. Upon activation, the system simultaneously triggers the audio output (TTS or Binaural Beat), modulates the matrix display to simulate a dynamic wave, and engages the servo motors for active face tracking.

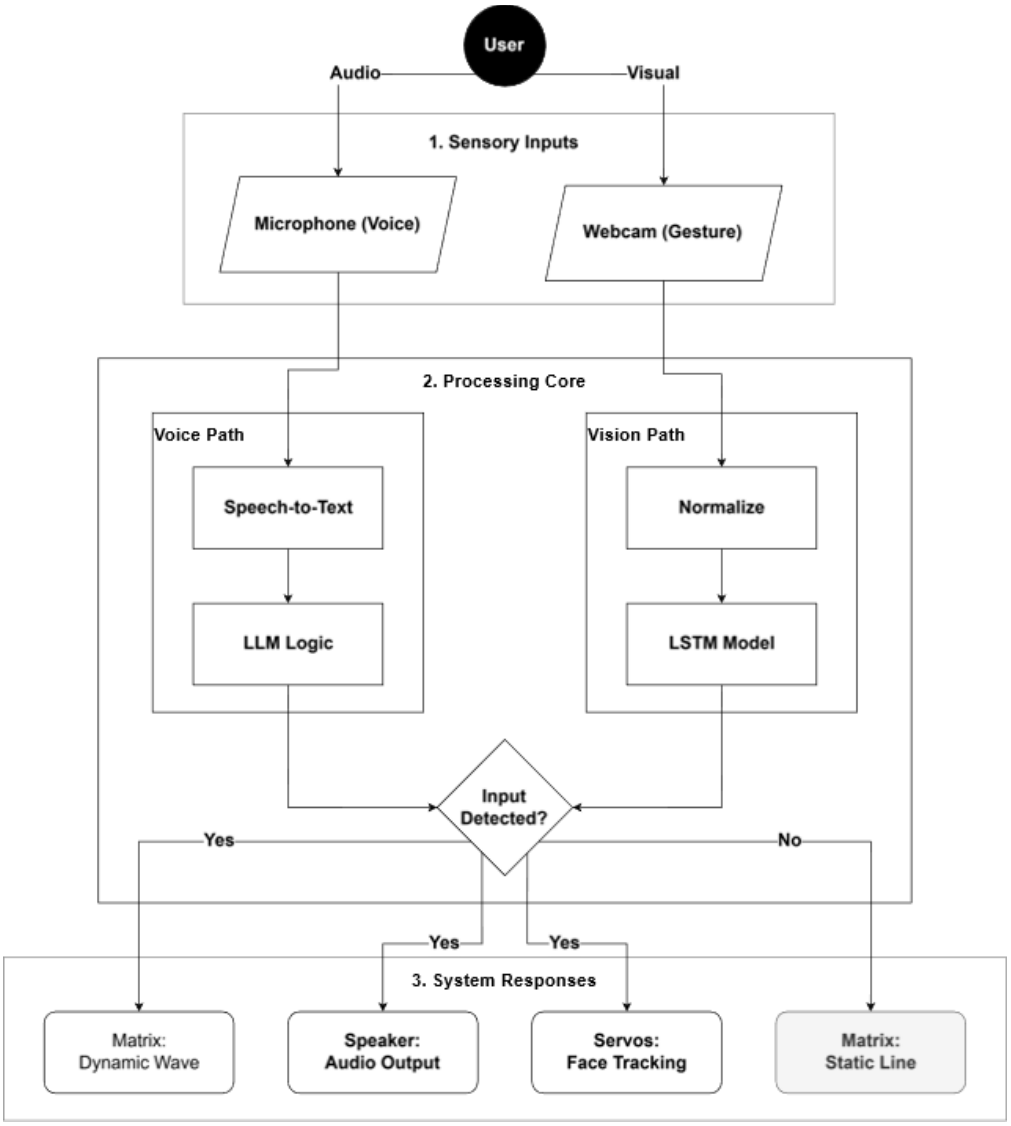

Figure 2: Operational flow of the multimodal system illustrating parallel processing pipelines for visual and auditory inputs.

The hardware architecture is orchestrated by the Raspberry Pi 4 Model B, which serves as the central logic controller for all peripheral coordination. To maintain system stability, the power distribution is split: a 10,000 mAh lithium-polymer unit powers the processor, while a separate supply feeds the actuators to prevent voltage drops during motor spikes. The coordination between the CPU and the mechanical components is managed through specific communication protocols. For the eye-tracking mechanism, the Raspberry Pi acts as the I2C master, sending coordinate data via the SDA/SCL lines to a PCA9685 driver. This driver offloads the Pulse Width Modulation (PWM) generation from the main CPU, ensuring stable control of the SG90 micro servos. Simultaneously, the L298N motor driver interfaces with the GPIO pins to manage the DC motors for base mobility, using an internal H-Bridge circuit to isolate high-current inductive loads from the sensitive logic board. For visual feedback, the MAX7219 dot matrix controller communicates via a serial interface. The Raspberry Pi synchronizes this display with the audio output state; when the Text-to-Speech engine is active, the Raspberry Pi modulates the matrix LEDs to simulate a dynamic wave, reverting to a static line during idle

ISSN 2394 - 7780

states. Complementing this visual feedback, the audio signals for speech and therapy are routed through the onboard 3.5mm jack to the speaker unit. On the input side, the high-speed USB interface handles the bandwidth-heavy video and audio streams from the webcam and microphone. This modular arrangement ensures that high-current actuation does not interfere with the logic circuits, maintaining consistent performance during simultaneous operation. Figure 3 details the precise wiring and signal paths between these components.

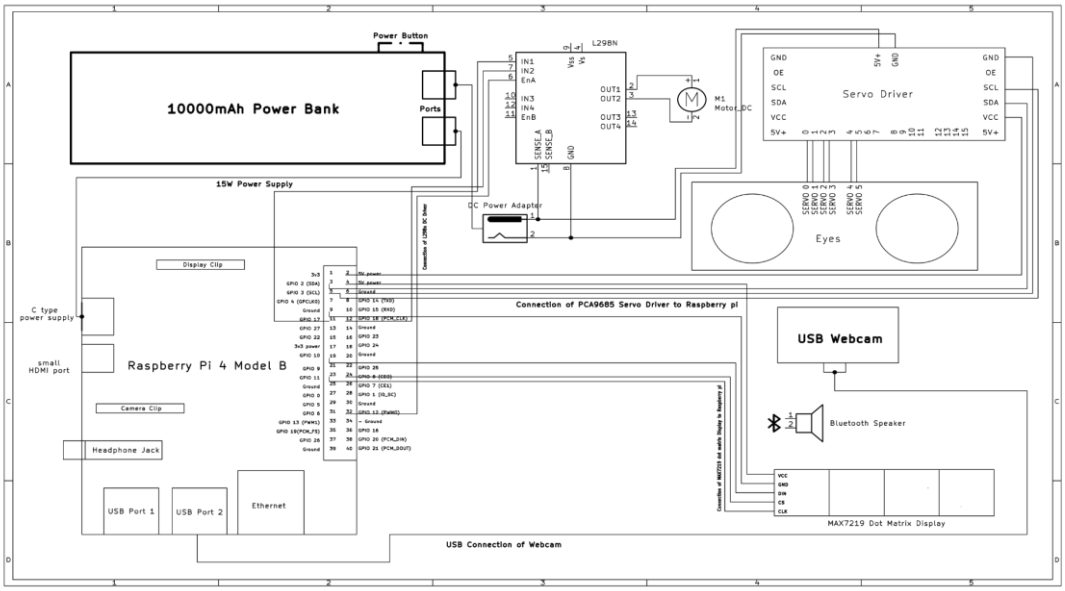

Figure 3: Circuit schematic detailing the interface between the Raspberry Pi 4, drivers, and sensory peripherals.

For the core algorithmic implementation, a LSTM network was selected over standard Recurrent Neural Networks (RNNs) to address the vanishing gradient problem inherent in processing long temporal sequences (Hochreiter & Schmidhuber, 1997). Furthermore, the LSTM architecture was prioritized over Gated Recurrent Units (GRUs) to leverage its distinct gating mechanism, which provides superior retention of long-term dependencies essential for distinguishing complex, multi-stage signs. The specific model architecture, detailed in Figure 4, utilizes a stacked configuration designed to extract spatial and temporal patterns. features. The network accepts an input tensor of shape (30×63), representing 30 consecutive frames of 21 vectorized landmarks. The first layer consists of 64 LSTM units with a configuration that retains the temporal context of the gesture. To prevent the model from memorizing training samples, a Dropout regularization layer is applied. This is followed by a second LSTM layer with 32 units to refine the feature extraction. Finally, the high-level features are passed to a Dense layer, and the classification is performed by a Softmax activation function.

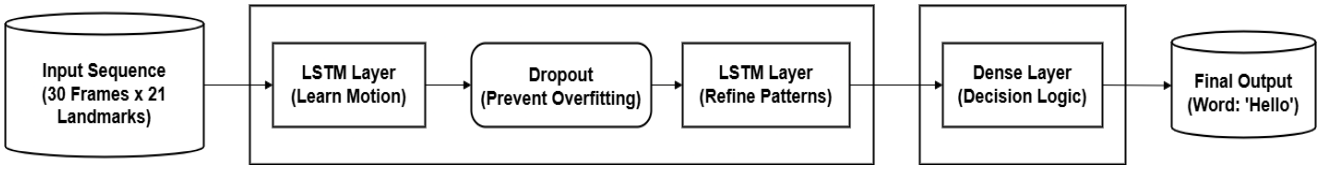

Figure 4: Architecture of the LSTM neural network designed to classify temporal sequences of MediaPipe landmarks.

Once the cognitive state was inferred, binaural beat parameters were generated algorithmically using a custom-designed adaptive frequency modulation algorithm. A binaural beat is produced by presenting two sinusoidal tones with slightly different frequencies to the left and right ears, resulting in the perception of a third beat frequency equal to the frequency difference between the two tones (Oster, 1973). A pure sinusoidal signal can be defined by the general formula for a sine wave (Kreyszig, 1979), as shown in Equation 3:

$$A(t) = sin(2\Pi ft) \quad (3)$$

where $A(t)$ represents the signal amplitude at time t, and f denotes the frequency in hertz. To generate the binaural effect, two signals were synthesized independently for each audio channel. The left channel signal was generated using Equation 4 and the right channel signal using Equation 5:

$$L(t) = sin(2\Pi f_{base}t) \quad (4)$$

ISSN 2394 - 7780

$$R(t) = sin(2\Pi(f_{base} + \Delta f)t) \quad (5)$$

Where $f_{base} = 200\ Hz$ is a fixed carrier frequency and Δf represents the beat frequency difference determined by the predicted cognitive state. The perceived binaural beat frequency is the absolute difference between the two channel frequencies (Hall, 1980). This adaptive mapping enables state-aware binaural beat generation aligned with EEG-derived cognitive states. It is important to note that the wellness module is designed for non-clinical relaxation support and does not replace medical intervention (Platt & Hammond, 2025).

The operational behavior of the proposed model "Zeus" is designed to create a natural and engaging interaction for the user. When the system detects a person, the robot's animatronic eyes automatically move to track their face, which establishes direct eye contact and makes the robot appear attentive. During interaction, such as when translating sign language into speech or playing therapeutic audio, the matrix display on the face animates with a moving wave pattern to simulate talking. If the robot is waiting or listening, this display changes to a simple static line. This combination of looking at the user and synchronizing the mouth lights with sound helps the robot feel like a helpful companion rather than just a machine. The final prototype was tested to ensure these physical reactions happen smoothly and in real-time. The assembled system used for this validation is shown in Figure 5.

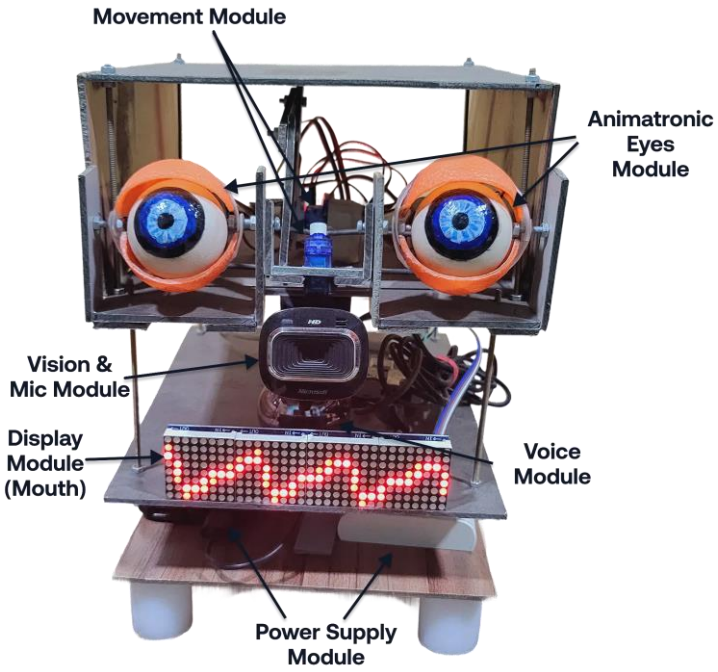

Figure 5: The assembled prototype of the proposed model "Zeus" configured for experimental validation.

4. RESULTS AND DISCUSSION

The proposed multimodal system was systematically evaluated to determine its efficacy in bridging the accessibility and wellness gaps. The experimental analysis focused on three primary performance metrics, including the classification accuracy of the ISL recognition module, the predictive precision of the wellness module, and the system latency of the distributed client-server architecture. The LSTM network was trained and validated using an 80-20 split on the Curated ISL Word Corpus, covering 69 dynamic gesture classes. As demonstrated in the experimental logs, the training loss converged effectively after 50 epochs, indicating stable learning without significant overfitting. The model achieved a final testing accuracy of 99.85% on the validation set. A granular analysis of the classification performance reveals high precision across both static and dynamic categories. For the static alphabet model (N = 1,380), validated against established data standards, representative classes such as 'A', 'B', and 'C' achieved perfect True Positive rates with zero False Positives, confirming the geometric validity of the feature pipeline (Wadhawan & Kumar, 2020). Similarly, the dynamic word model demonstrated robustness; specific gestures like 'Thank You' and 'Hello' recorded high True Positive counts with negligible misclassifications, confirming the model's ability to distinguish temporal features effectively.

ISSN 2394 - 7780

Table 2. Comparative Analysis of SLR Architectures

Author	Methodology	Accuracy	System Latency / Type
Jiang et al. (2021)	Skeleton Graph (SAM-SLR)	98.42%	High (Heavy GPU required)
Katoch et al. (2022)	CNN+SVM (Static)	99.64%	High (offline/Static)
Subramanian et al. (2022)	MediaPipe + GRU	95.00%	Low (Real-Time Edge)
Meng et al. (2024)	MediaPipe + FingerNet	95.30%	Low (Real-Time)
Trisha et al. (2025)	Contour Detection	90.00%	Medium (~ 0.8 s)
Proposed System	MediaPipe + LSTM (Distributed)	99.85%	Low (< 200 ms)

A comparative analysis, detailed in Table 2, illustrates the performance trade-offs present in existing techniques. High accuracy using CNNs has been documented, though such methods were focused on static image classification (Katoch et al., 2022). Conversely, lightweight edge solutions successfully optimized for computational efficiency on resource-constrained hardware, resulting in varied accuracy rates for dynamic gestures (Subramanian et al., 2022; Trisha et al., 2025). The proposed method seeks to balance these parameters, achieving 99.85% accuracy while maintaining an average end-to-end latency of 180 ms. This performance successfully meets the real-time interaction benchmark of < 200 ms established in the research hypothesis. This balance is attributed to the distributed client-server design, which leverages the high processing power of the server for the LSTM network while maintaining the portability of the edge device.

Furthermore, the system successfully validates the role of affective computing in enhancing robotic performance. The wellness module, utilizing the Random Forest classifier, was evaluated on the test subset of the EEG Emotion Dataset (Mohsen, 2021). It achieved a classification accuracy of 91.7% in distinguishing between focused, relaxed, and stressed states. Performance analysis indicates a strong ability to identify critical states; specifically, the 'Stressed' class yielded 31 True Positives and zero False Negatives, ensuring that users requiring therapeutic intervention were reliably detected. The 'Focused' and 'Relaxed' states similarly showed high recognition rates, validating the reliability of the trigger mechanism for binaural beats. Additionally, the integration of this wellness module introduces a layer of affective intelligence to the SLR framework, confirming that the system can reliably differentiate between cognitive states to deploy the correct therapeutic intervention.

5. CONCLUSION

The primary objective of this research was to engineer a cost-effective, multimodal system capable of bridging the accessibility and wellness gaps in social robotics. The experimental findings explicitly confirm the validity of the proposed solution against the initial research hypotheses. The hypothesis regarding inclusivity was validated as the MediaPipe-optimized LSTM model achieved a testing accuracy of 99.85% on the Curated ISL Word Corpus, surpassing the 90% success threshold defined in the research design. Similarly, the hypothesis concerning real-time performance was confirmed; the distributed client-server architecture maintained an average system latency of 180 ms, successfully meeting the sub-200 ms benchmark required for fluid human-robot interaction. Furthermore, the integration of affective computing proved effective, as the wellness module, powered by a Random Forest classifier trained on the EEG Human Emotion Dataset, demonstrated an efficacy rate of 91.7% in identifying cognitive states. This validates the system's ability to accurately deploy therapeutic 10 Hz Alpha binaural beats based on user needs. Collectively, these results prove that complex embodied AI does not require industrial-grade hardware, establishing a scalable foundation for future developments in inclusive home automation and personalized robotic companionship.

RECOMMENDATIONS

The findings revealed that while the current distributed architecture ensures high performance, it offers opportunities for significant functional expansion. Future work will focus on integrating Smart Home Automation (IoT) protocols, allowing the robot to control environmental factors (lights, temperature) to further aid the DHH community. We also aim to expand the vision pipeline to recognize continuous sign language sentences rather than isolated gestures. Additionally, future iterations will enhance the system's Personalized AI capabilities by fine-tuning the Large Language Model (LLM) on user-specific interactions, creating a more adaptive and context-aware companion.

Acknowledgement

The authors express their gratitude to the Department of Computer Science at Sheth L.U.J. & Sir M.V. College of Arts, Science & Commerce for their academic support and encouragement. We extend our sincere

ISSN 2394 - 7780

appreciation to our project supervisor, Dr. Mahendra Kanojia, for his valuable insights, technical guidance, and mentorship throughout the development of this research.

Funding Support
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The project was entirely self-funded by the authors, who contributed equally to the procurement of hardware components and development resources.

Ethical Statement
This study primarily focuses on robotic system design and software architecture. All participation in the system testing phase was voluntary, and informed consent was obtained from all individuals involved. No clinical trials or invasive medical experiments were conducted.

Conflicts of Interest
The authors declare that they have no conflicts of interest regarding the publication of this paper.

Data Availability Statement
Data is available on request from the corresponding author upon reasonable request.

REFERENCES
Al-Shargie, F., Tang, T. B., & Kiguchi, M. (2022). Stress management using fNIRS and binaural beats stimulation. *Biomedical Optics Express*, 13(5), 2686–2704. https://pmc.ncbi.nlm.nih.gov/articles/PMC9208616/

Bird, J. J., Ekárt, A., Buckingham, C. D., & Faria, D. R. (2019). Mental emotional sentiment classification with an EEG-based brain-machine interface. *The International Conference on Digital Image and Signal Processing (DISP'19)*. https://www.researchgate.net/publication/329403546_Mental_Emotional_Sentiment_Classification_with_an_EEG-based_Brain-machine_Interface

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 125. https://doi.org/10.3390/info11020125

Hall, G. L. (1980). *The hearing aid: Its operation and development*. National Hearing Aid Society.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hong, A., et al. (2021). A multimodal emotional human–robot interaction architecture for social robots engaged in bidirectional communication. *IEEE Transactions on Cybernetics*, 51(12), 5954–5968. https://ieeexplore.ieee.org/abstract/document/9655474

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. https://ieeexplore.ieee.org/document/9523142/

Jurvanen, K. (2020). *Binaural beats and music: Using theta and alpha waves in music to induce relaxation* [Master's thesis, Aalto University].

Karthik, S., Deshmukh, S., Save, A., & Shah, R. (2024). Effective communication between blind, mute and deaf people using a multi-model approach. *International Research Journal of Engineering and Technology (IRJET)*, 11(3), 1518–1522.

Katoch, S., Singh, V., & Tiwary, U. S. (2022). Indian Sign Language recognition system using SURF with SVM and CNN. *Array*, 14, 100141. https://linkinghub.elsevier.com/retrieve/pii/S2590005622000121

Kreyszig, E. (1979). *Advanced engineering mathematics* (4th ed.). Wiley.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*. http://arxiv.org/abs/2005.11401

Lugaresi, C., et al. (2019). MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*. https://arxiv.org/abs/1906.08172

Ma, Y., Zhang, Y., Bachinski, M., & Fjeld, M. (2023). Emotion-aware voice assistants: Design, implementation, and preliminary insights. *Proceedings of the 11th International Symposium on Chinese CHI*, 527–532. https://dl.acm.org/doi/fullHtml/10.1145/3629606.3629665

Masina, F., Orso, V., Pluchino, P., Dainese, G., Volpato, S., Nelini, C., Mapelli, D., Spagnolli, A., & Gamberini, L. (2020). Investigating the accessibility of voice assistants with impaired users: Mixed methods study. *Journal of Medical Internet Research*, 22(9), e18431. http://www.jmir.org/2020/9/e18431/

Meng, Y., Jiang, H., Duan, N., & Wen, H. (2024). Real-time hand gesture monitoring model based on MediaPipe's registerable system. *Sensors*, 24(19), 6262. https://www.mdpi.com/1424-8220/24/19/6262

Mohsen, S. (2021). EEG human emotion dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/drsaeedmohsen/eeghumanemotiondataset2021

Oster, G. (1973). Auditory beats in the brain. *Scientific American*, 229(4), 94–102. https://doi.org/10.1038/scientificamerican1073-94

Pan, M. K. X. J., & Law, E. (2020). Realistic and interactive robot gaze. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11050–11060. https://ieeexplore.ieee.org/document/9341297/

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *IARJSET*, 2(3), 20–22. https://doi.org/10.17148/IARJSET.2015.2305

Platt, J., & Hammond, L. (2025). Is non-clinical, personal use of binaural beats audio an effective stress-management strategy? A systematic review of randomised control trials. *Advances in Mental Health*, 23(2), 258–286. https://www.tandfonline.com/doi/full/10.1080/18387357.2024.2374759

Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175(4), 7–9. https://doi.org/10.5120/ijca2017915495

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. https://doi.org/10.1037/h0077714

Shalforoushan, S. M., & Golmakani, Z. B. (2022). The effectiveness of alpha binaural beats in reducing stress and rumination and promoting sleep quality in university students. *Journal of Sleep Sciences*, 6(3-4), 67–73. https://publish.kne-publishing.com/index.php/JSS/article/view/10885

Sta Maria, T. J., & Deja, J. A. (2024). Alexa, I wanna see you: Envisioning smart home assistants for the deaf and hard-of-hearing. *arXiv*. https://arxiv.org/pdf/2412.00514.pdf

Subramanian, B., Olimov, B., Naik, S. M., Kim, S., Park, K. H., & Kim, J. (2022). An integrated mediapipe-optimized GRU model for Indian sign language recognition. *Scientific Reports*, 12, 11964. https://www.nature.com/articles/s41598-022-15998-7

Trisha, K., Varma, K. T. A., Ramu, K., Sekhar, K. S., Sampath, L., Raju, N., & Prasad, B. S. (2025). Raspberry Pi and OpenCV based sign language recognition system for mute community. *NSRIT Journal of Electronics and Communication Engineering*.

Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, 32(12), 7957–7968. https://doi.org/10.1007/s00521-019-04691-y

World Health Organization. (2021). *World report on hearing*. https://www.who.int/publications/i/item/world-report-on-hearing

Xiong, J., et al. (2025). Binaural beats for perioperative anxiety and pain: A systematic review and meta-analysis. *Complementary Therapies in Medicine*, 95, 103299. https://linkinghub.elsevier.com/retrieve/pii/S096522992500175X

Zuo, R., Wei, F., & Mak, B. (2024). Towards online continuous sign language recognition and translation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11050–11060. https://aclanthology.org/2024.emnlp-main.619