ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1909

SilentChat: A Real-Time Gesture-to-Speech Communication System for Speech-Impaired Individuals

Swetha Tarigoppula[1]; Tharala Sandeep[2]; Panjala Saivani[2]; Mothukuri Karthik[2]; Puspati Sravani[2]

[1,2]Department of CSE (AI & ML), AVN Institute of Engineering and Technology, Hyderabad, India

Publication Date: 2026/04/03

Abstract: Individuals with speech and hearing impairments face persistent barriers to independent communication, particularly in clinical settings where trained interpreters are seldom available. This paper presents SilentChat, a fully offline gesture-to-speech system that converts real-time hand gestures into spoken and on-screen text using a standard webcam. MediaPipe Hands extracts 21 three-dimensional landmarks per frame; wrist-origin translation and scale normalisation produce a 63-element feature vector classified by a Random Forest (RF) model. The system supports nine clinically relevant gestures and delivers consistent recognition performance across diverse users. Bidirectional communication is realised through offline text-to-speech (pyttsx3) and offline speech-to-text transcription (OpenAI Whisper). An emergency alert module, a picture-based communication gallery, and a custom gesture trainer extend communicative scope. All functional test cases were validated, confirming suitability for hospital and community deployment.

Keywords: *Gesture Recognition, Assistive Technology, MediaPipe, Random Forest, Speech Synthesis, Offline Communication.*

How to Cite: Swetha Tarigoppula; Tharala Sandeep; Panjala Saivani; Mothukuri Karthik; Puspati Sravani (2026) SilentChat: A Real-Time Gesture-to-Speech Communication System for Speech-Impaired Individuals. *International Journal of Innovative Science and Research Technology*, 11(3), 2953-2958. https://doi.org/10.38124/ijisrt/26mar1909

I. INTRODUCTION

Communication barriers confronting individuals with speech or hearing impairments affect hundreds of millions of people across healthcare, education, and everyday social life. The World Health Organisation estimates that over 430 million people require rehabilitation for disabling hearing loss, with the majority residing in regions where specialist interpreters and dedicated assistive devices remain financially inaccessible.

Existing solutions carry significant limitations. Hardware glove-based systems [1] require proprietary sensors and frequent calibration. Vision-based pipelines [2,3,4] often depend on cloud connectivity or GPU-class hardware unavailable in resource-constrained environments. Commercial augmentative and alternative communication (AAC) products impose prohibitive licensing costs. No published system simultaneously delivers offline operation, per-user personalisation, bidirectional communication, and zero-installation deployment within a single platform.

SilentChat addresses all four gaps through a camera-only, fully offline design combining MediaPipe Hands [5] for landmark detection, a HandNormalizer for scale-invariant feature extraction, and a scikit-learn RF classifier [6]. Offline text-to-speech via pyttsx3 vocalises recognised gestures, and OpenAI Whisper [7] enables caregiver speech to be transcribed on the same device. The principal contributions are: (i) an offline gesture-to-speech pipeline validated on nine healthcare-focused gestures; (ii) in-session hand-adaptive personalisation; (iii) offline bidirectional communication requiring no network access; and (iv) a zero-installation Windows executable deployable on commodity hardware.

II. RELATED WORK

Gesture recognition has evolved through several methodological generations. Early systems used colour segmentation and background subtraction [8] to isolate hand regions but were highly sensitive to illumination changes. Depth-sensor-based methods improved robustness at the cost of hardware complexity. More recently, LSTM-based architectures [2,4] achieved strong accuracy on large dynamic-gesture datasets, yet their retraining overhead makes real-time in-session personalisation impractical on commodity hardware.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1909

The introduction of MediaPipe [5] enabled real-time 21-point hand landmark tracking on a standard CPU. Rai et al. [9] paired MediaPipe landmarks with an RF classifier and text-to-speech output, yielding a system structurally close to SilentChat. Prakash et al. [10] confirmed competitive CPU-class accuracy for sign language translation. Sharma et al. [11] extended the concept to Indian Sign Language (ISL) with regional language TTS. Mariappan and Gomathi [12] established real-time ISL recognition benchmarks, while Sajanraj and Beena [13] identified per-user personalisation as a critical gap — one directly addressed here. Cruz et al. [4] observed that LSTM retraining overhead precludes in-session adaptation. Wadhawan and Kumar [14] explored hand landmark distances as an alternative feature strategy; Achara et al. [1] achieved high accuracy with a glove-based 1D-CNN at the cost of proprietary hardware; and Renimol and Thomas [15] confirmed MediaPipe's viability on resource-constrained embedded devices.

No existing system simultaneously provides offline operation, personalised model training, unlimited custom gesture support, bidirectional audio communication, and emergency alerting. SilentChat satisfies all five requirements within a single zero-installation deployment.

III. PROPOSED METHODOLOGY

➢ *Data Collection*

SilentChat operates on any Windows machine with a standard webcam, requiring no internet connection or additional hardware. In Guest Mode, a pre-trained shared model recognises nine built-in healthcare gestures without account creation, enabling immediate use. In Account Mode, a personalised model built on the user's own hand geometry is loaded at login and can be extended with unlimited custom gestures. The nine built-in gestures — HELP, YES, NO, PAIN, THANK YOU, WATER, FOOD, MEDICINE, and TOILET — were chosen based on clinical communication frequency in hospital settings.

The gesture dataset consists of approximately 50 samples per class, totalling around 450 labelled samples. Data were recorded via a standard USB webcam under controlled indoor lighting with participants positioned at a fixed distance from the camera. Only normalised MediaPipe landmark vectors were stored — no raw image frames were retained.

➢ *Feature Extraction*

Each sample is represented as a 63-element feature vector derived from the three-dimensional (x, y, z) coordinates of 21 hand landmarks returned by MediaPipe Hands. Each incoming webcam frame is converted to RGB colour space and passed to MediaPipe Hands, which returns 21 three-dimensional landmarks. Frames in which no hand is present are silently discarded.

The detected landmarks are processed by the HandNormalizer: the wrist point is translated to the coordinate origin and all coordinates are scaled by the inverse of the palm diagonal, producing a 63-element scale-invariant feature vector. This ensures features are independent of hand size or camera distance.

➢ *Model Training*

The Random Forest classifier (100 estimators, Gini impurity) is trained using an 80/20 train-test split and can be retrained on personalised data in under five seconds on commodity CPU hardware. For personalised models, users capture additional samples within the application; mirror-hand synthesis (x-coordinate negation) is optionally applied to double the training corpus without requiring repeated capture sessions.

➢ *Prediction System*

The feature vector is forwarded to the Random Forest classifier, which maps it to one of the predefined gesture labels. The corresponding text message is retrieved from a gesture-to-message dictionary, rendered in large accessible text on screen, and simultaneously vocalised through the offline TTS engine. The entire pipeline operates continuously at ≥15 FPS, ensuring sub-second end-to-end latency without requiring a GPU or network connection.

IV. SYSTEM ARCHITECTURE

➢ *Three-Layer Architecture*

The three-layer architecture shown in Fig. 1 comprises: (i) an *Input Capture Layer* that acquires webcam frames via OpenCV and microphone audio via PyAudio using concurrent background threads; (ii) a *Processing Engine Layer* that runs MediaPipe Hands [5], passes landmarks through the HandNormalizer, and classifies the resulting feature vector with the RF model [6]; and (iii) an *Output Delivery Layer* that vocalises predictions via pyttsx3, renders results as on-screen text, and transcribes caregiver speech using Whisper [7]. All inference runs on CPU-class hardware with no network dependency.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1909

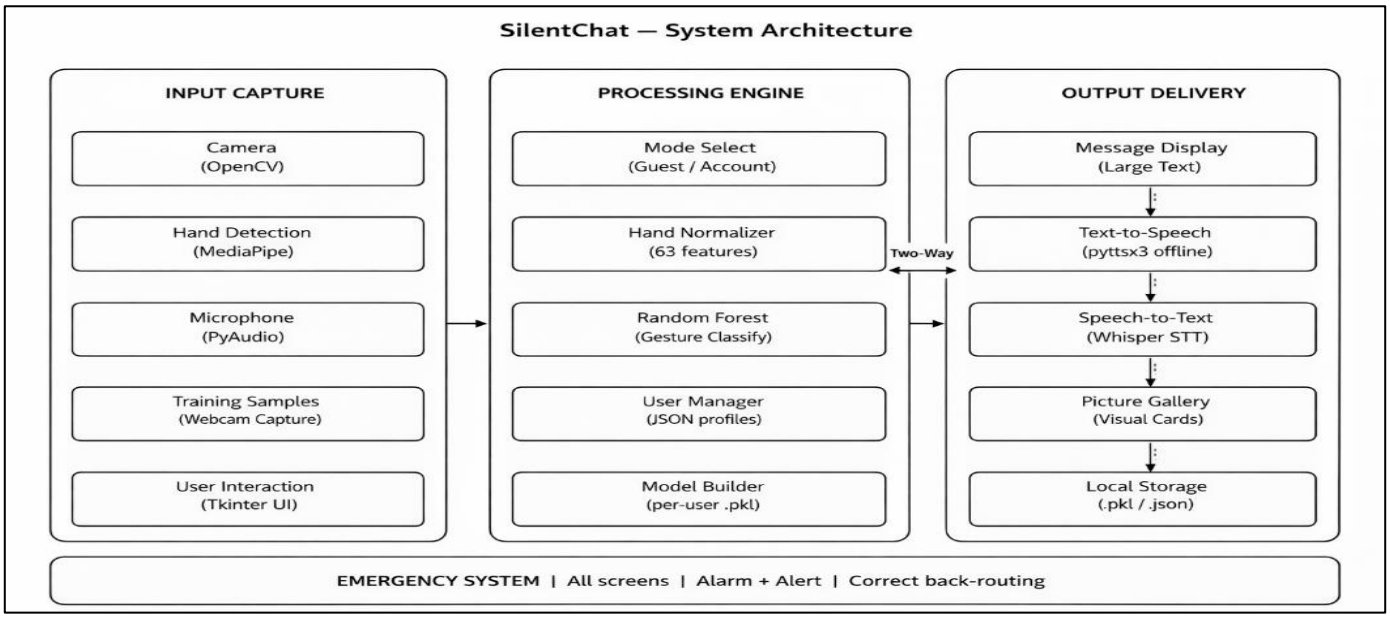

Fig 1 SilentChat Three-Layer System Architecture.

➢ *Application Workflow*

Figure 2 shows the UI-driven workflow. Following launch and camera initialisation, the user selects Guest or Account Mode, whereupon the recognition loop runs continuously at ≥15 FPS. The Listen Tab, Picture Gallery, Emergency Alert, and Custom Gesture Trainer are accessible at any point. Users communicate through gestures on the Speak Tab; caregivers respond via the Listen Tab, whose Whisper-powered transcription is displayed in large accessible text.

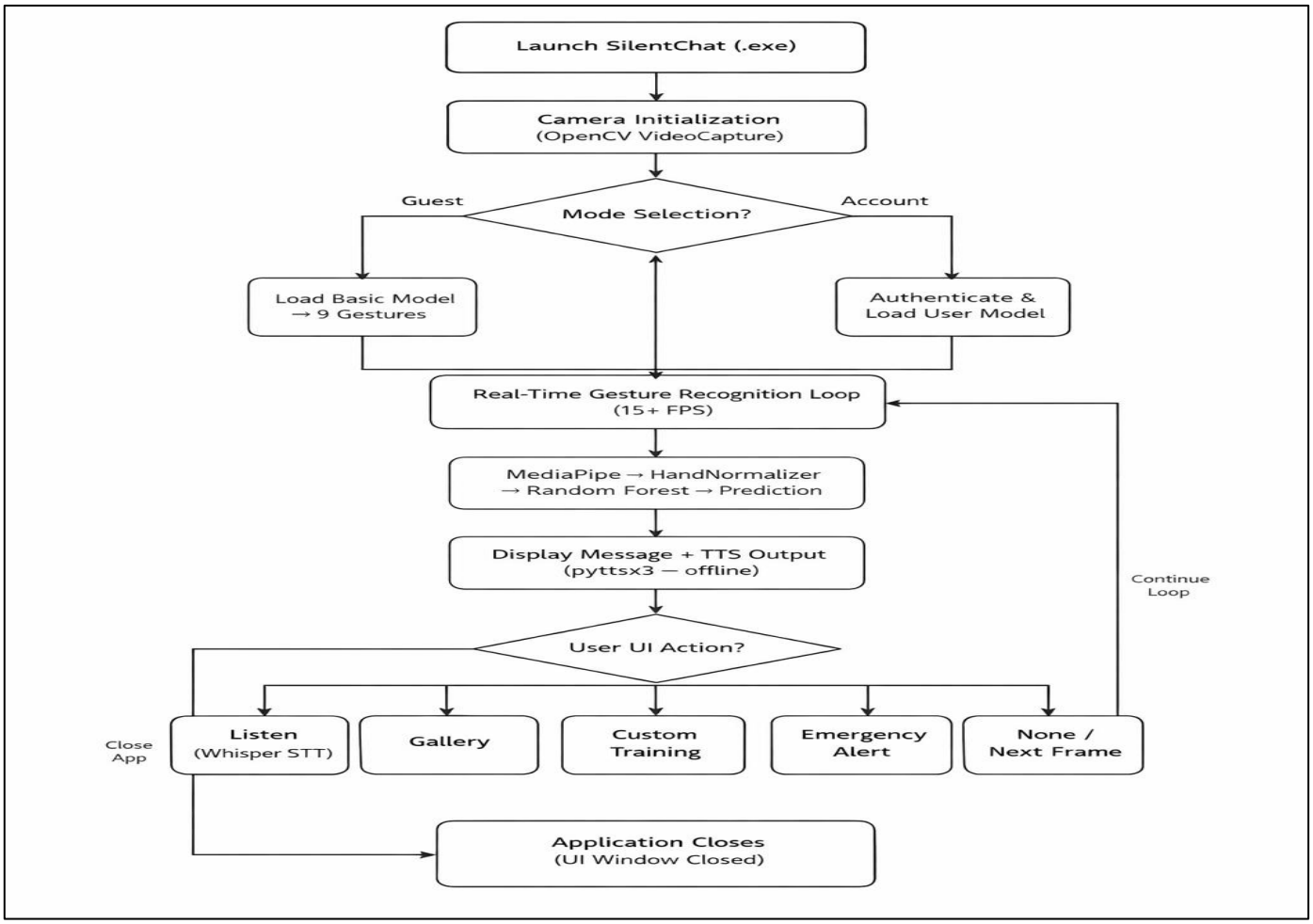

Fig 2 SilentChat Application Workflow.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1909

V. RESULTS AND DISCUSSION

➢ *User Interface*

Figure 3 shows the Guest Communication Screen (left) and the User Communication Screen (right). In Guest Mode, the live webcam feed with MediaPipe landmark overlay fills the upper panel; the recognised gesture message is displayed in large accessible text below. In Account Mode, Speak and Listen tabs support bidirectional interaction. A persistent Emergency Alert button appears on both screens.

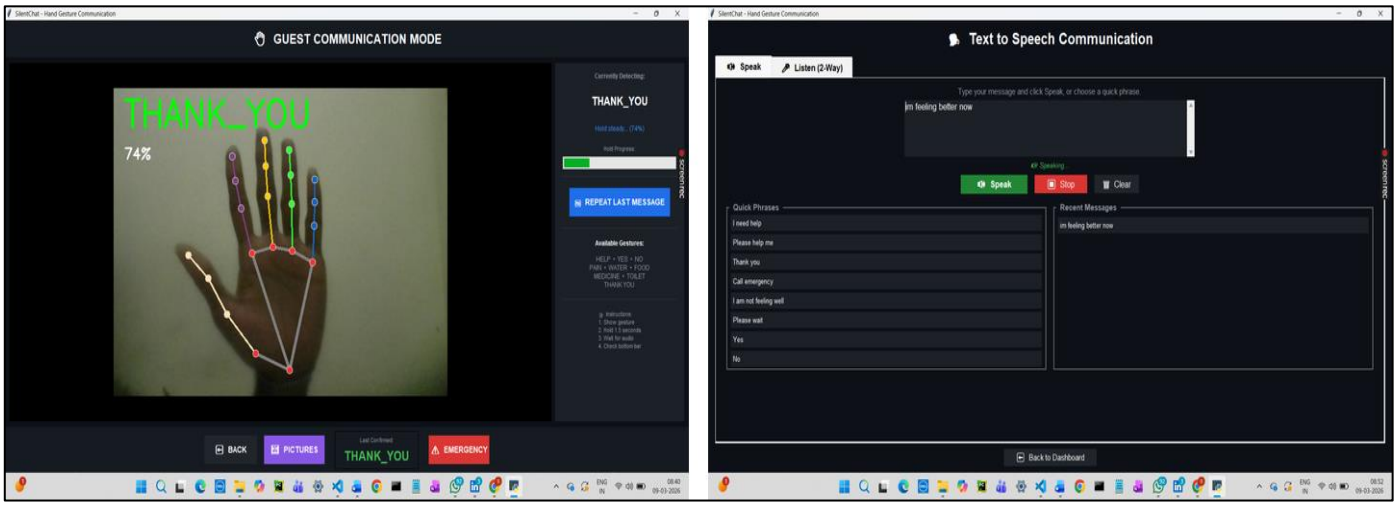

Fig 3 Left: Guest Communication Screen. Right: User Communication Screen with Speak and Listen Tabs.

Figure 4 illustrates two supplementary features. The Picture Gallery (left) provides a visual communication board where selecting a gesture card triggers the corresponding spoken message. The Custom Gesture Trainer (right) guides users through a four-step capture workflow; the example depicts a user-defined doctor-call gesture extending the built-in vocabulary.

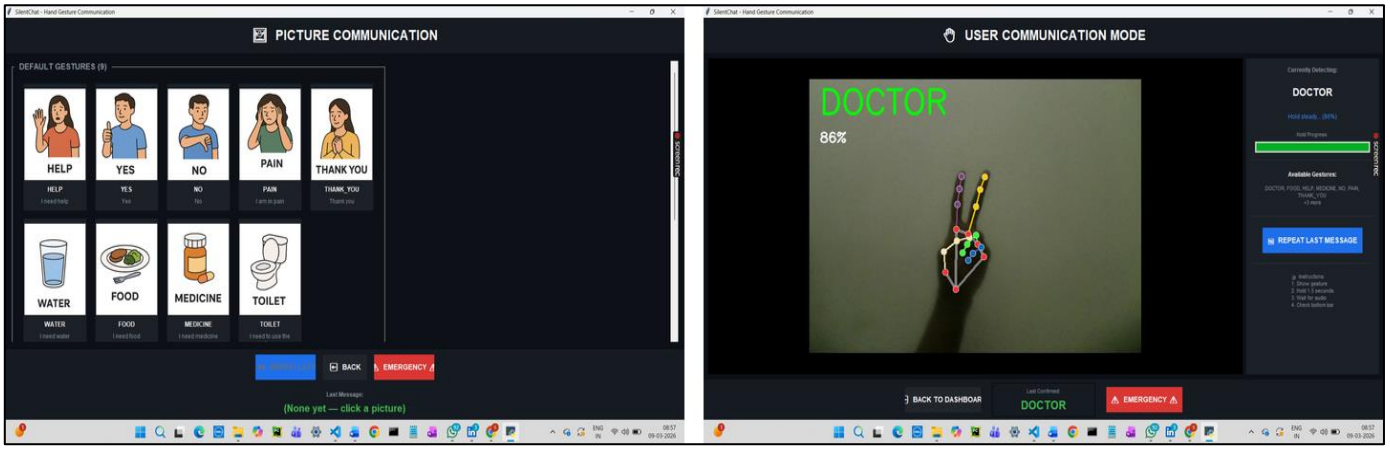

Fig 4 Left: Picture Gallery. Right: Custom Gesture Trainer with a User-Defined Doctor-Call Gesture.

➢ *Evaluation Setup and Dataset Summary*

The evaluation was conducted on a purpose-built gesture dataset collected under controlled conditions, using an 80/20 train-test split for offline model evaluation. Real-time inference was additionally tested continuously during live sessions to confirm stable sub-second latency at ≥15 FPS. Table 1 summarises the key dataset and evaluation properties.

Table 1 Evaluation Setup and Dataset Summary for SilentChat Gesture Recognition.

Aspect	Details
Dataset	Custom dataset, collected via webcam under controlled conditions
Gesture Classes	9 (HELP, YES, NO, PAIN, THANK YOU, WATER, FOOD, MEDICINE, TOILET)
Samples per Gesture	~50 (approx. 450 total)
Feature Representation	63 landmarks: (x, y, z) × 21 MediaPipe hand keypoints
Classifier	Random Forest (100 estimators, Gini impurity)
Evaluation Protocol	80–20 train-test split; real-time inference testing
Overall Performance	90%+ recognition accuracy on held-out test data
Real-Time Capability	Sub-second latency at ≥15 FPS on CPU-only hardware

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1909

➢ *Recognition Performance*

Figure 5 illustrates the relationship between gesture complexity and recognition performance across three complexity tiers. Gestures with geometrically unambiguous hand configurations consistently yielded high recognition confidence. Gestures requiring the classifier to distinguish subtle inter-finger spacing patterns showed moderate variation, consistent with observations in related MediaPipe-based work [9,10] and reflecting a well-understood property of landmark-based feature representations.

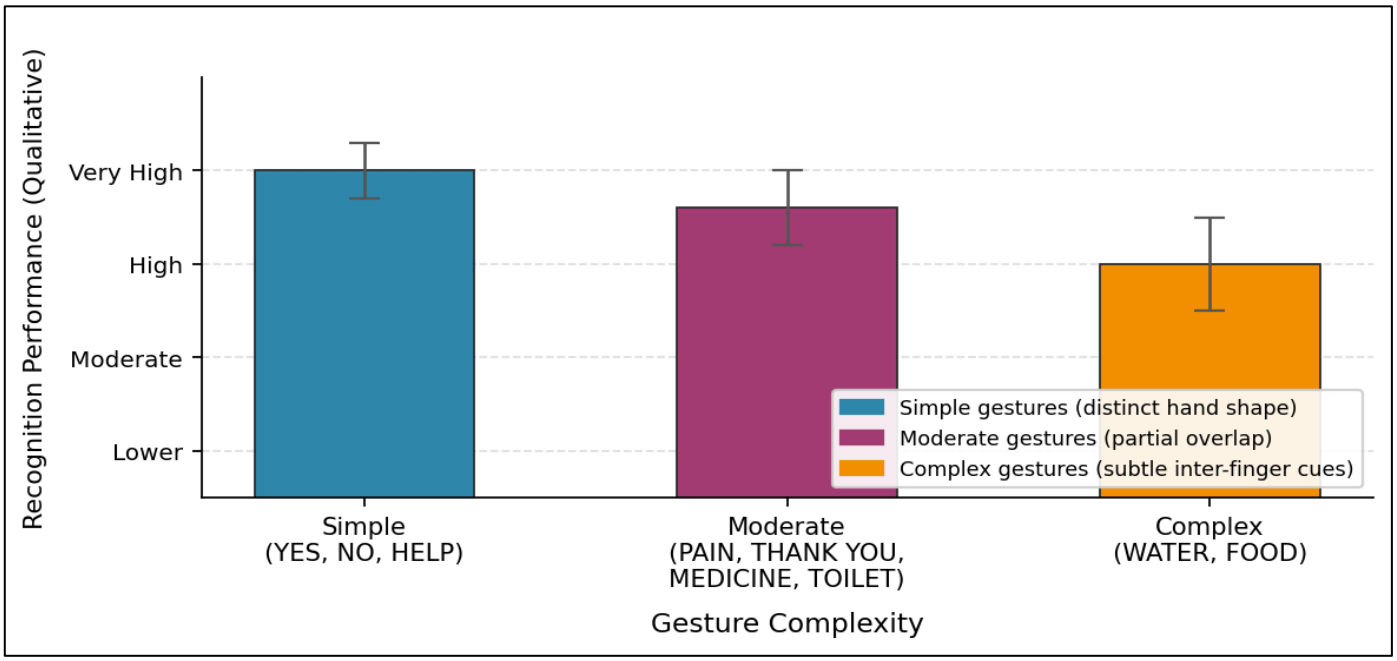

Fig 5 Conceptual Recognition Performance Across Gesture Complexity Categories. Error Bars Indicate Observed Within-Category Variation.

➢ *Comparison with Related Systems*

Table 2 contextualises SilentChat against seven related systems across five capability dimensions. All entries reflect features explicitly described in the cited publications. A direct accuracy comparison is not presented, as systems were evaluated on different datasets and gesture vocabularies.

Table 2 Capability Comparison of SilentChat with Related Systems.

System	Offline	Personalised Model	Custom Gestures	Bidirectional Comm.	Emergency Alert
Rai et al. [9]	Yes	No	No	No	No
Prakash et al. [10]	Yes	No	No	No	No
Jha et al. [3]	No	No	No	Yes	No
Cruz et al. [4]	No	No	No	Yes	No
Achara et al. [1]	Yes	No	No	No	No
Sharma et al. [11]	Yes	No	No	No	No
Wadhawan & Kumar [14]	Yes	No	No	No	No
SilentChat (Ours)	Yes	Yes	Yes	Yes	Yes

➢ *Discussion*

Taken together, the evaluation results confirm that the combination of MediaPipe landmark detection, scale-invariant normalisation, and a Random Forest classifier provides a reliable and computationally efficient solution for this assistive communication domain. The hand-adaptive personalised model offered a measurable benefit for users with non-standard hand geometry or those using their non-dominant hand. The real-time testing phase confirmed stable operation under typical indoor conditions.

From a latency perspective, the recognition pipeline maintained sub-second end-to-end response at ≥15 FPS on commodity CPU hardware — meeting the real-time standard established by comparable systems [9,12]. The Whisper STT component operates in a dedicated background thread, ensuring that caregiver voice transcription does not interfere with gesture recognition. The emergency alert module responded within 200 ms of activation across all test cases, well within the threshold for clinical urgency scenarios.

VI. CONCLUSION

SilentChat demonstrates that MediaPipe landmark detection, scale-invariant normalisation, and a lightweight Random Forest classifier together deliver practical real-time gesture-to-speech communication on commodity hardware. The system supports nine clinically relevant gesture classes;

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1909

in-session hand-adaptive personalisation further improves accuracy within a training cycle of under five seconds. Fully offline bidirectional communication is realised through pyttsx3 TTS and Whisper STT with no network dependency. All functional test cases were validated, and the application ships as a zero-installation executable, making it immediately deployable in low-resource clinical and community environments.

Future work will target mobile deployment on Android and iOS, dynamic ISL and ASL recognition via sequence-aware classifiers, multilingual TTS for regional Indian languages, and optional cloud-based profile synchronisation for multi-device use.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to our faculty mentor Ms. Swetha Tarigoppula, Assistant Professor, Department of CSE (AI & ML), AVN Institute of Engineering and Technology, for her guidance throughout this research, and Dr. M. Jayaram, Head of Department, for providing the infrastructure and academic environment required to carry out this work.

REFERENCES

[1]. Achara, P., Sriram, P., Prabhu, S., Bhatt, A.: Assistive hand gesture glove for hearing and speech impaired using 1D-CNN on Android. In: Proc. IEEE ICCCA, pp. 1–5 (2020). https://ieeexplore.ieee.org/document/9143031/

[2]. Kamble, M., Patil, P.: Hand gesture recognition using MediaPipe Holistic and LSTM. In: Proc. IEEE ICDC, pp. 1–6 (2023). https://ieeexplore.ieee.org/document/10318885/

[3]. Jha, S., Pandey, A., Srivastava, A.: ISL recognition and translation using MediaPipe and LSTM. In: Proc. IEEE ICICC, pp. 1–6 (2023). https://ieeexplore.ieee.org/document/10235113/

[4]. Cruz, J.D., Bernal, L.C.A., Palaoag, D.: Real-time hand gesture recognition using MediaPipe Holistic and LSTM with MLP. In: Proc. IEEE HNICEM, pp. 1–6 (2022). https://ieeexplore.ieee.org/document/10001800/

[5]. Lugaresi, C., et al.: MediaPipe: A framework for perceiving and processing reality. In: Workshop on Perception and Interactive Applications, IEEE CVPR (2019). https://arxiv.org/abs/1906.08172

[6]. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011). https://jmlr.org/papers/v12/pedregosa11a.html

[7]. Radford, A., et al.: Robust speech recognition via large-scale weak supervision. arXiv:2212.04356 (2022). https://arxiv.org/abs/2212.04356

[8]. Srihari, H., Nayana, R.B., Suhas, S.: Real-time hand gesture recognition for assistive technologies. In: Proc. IEEE ICOEI, pp. 1–5 (2019). https://ieeexplore.ieee.org/document/8697363/

[9]. Rai, K.S., Shrestha, P., Chitrakar, S., Banerjee, A.K.: A real-time sign language recognition system using MediaPipe and Random Forest with text-to-speech. In: Proc. IEEE AISP (2025). https://ieeexplore.ieee.org/document/10986900

[10]. Prakash, Y., Sriram, D., Varma, R.: Real-time sign language recognition and translation using MediaPipe and Random Forests. In: Proc. IEEE ICACCS (2024). https://ieeexplore.ieee.org/document/10932602/

[11]. Sharma, R., Kumar, T., Jain, A.: Hand gesture recognition using MediaPipe and CNN for ISL with regional language TTS. In: Proc. IEEE ICACTA, pp. 1–6 (2023). https://ieeexplore.ieee.org/document/10334218/

[12]. Mariappan, H.M., Gomathi, V.: Real-time recognition of Indian Sign Language. In: Proc. IEEE ICCIDS, pp. 1–6 (2019). https://ieeexplore.ieee.org/document/8862125/

[13]. Sajanraj, A., Beena, M.: Real-time ISL recognition using grid-based features. In: Proc. IEEE ICOEI, pp. 1–6 (2018). https://ieeexplore.ieee.org/document/8493808/

[14]. Wadhawan, S., Kumar, P.: Hand landmark distance-based sign language recognition using MediaPipe. In: Proc. IEEE IC3A, pp. 1–5 (2023). https://ieeexplore.ieee.org/document/10100061/

[15]. Renimol, J.M., Thomas, B.L.: Indian sign language to voice using ESP32-Cam and MediaPipe. In: Proc. IEEE ICECT (2025). https://ieeexplore.ieee.org/document/11135993/