IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 4, December 2025

Impact Factor: 7.67

SignSpeak AI: Converting Sign Language to Text and Speech

Nandini S[1], Shuchitha G[2], Sai Deekshitha B P[3], Mehak Azmath[4], Mounika H J[5]

Associate Professor, Dept. of Information Science and Engineering [1]

Student, Dept. of Information Science and Engineering[2345]

SJC Institute of Technology, Chickballapur, Karnataka, India

Abstract: *SignSpeak AI is a real-time assistive framework designed to break down communication barriers for the hearing- impaired community. The system utilises a fine-tuned YOLOv8 deep learning model to translate static Indian Sign Language (ISL) finger-spelling gestures into meaningful text and synthesised speech. The model was trained on a custom- curated dataset spanning 26 alphabetical classes, rigorously stratified into a 70% training, 15% validation, and 15% testing split to ensure robust generalisation. Performance evaluations validate the system's high fidelity, with optimal gesture classes achieving a Precision of 0.949, a Recall of 1.0, and a mean Average Precision (mAP50) of 0.992. Beyond core recognition, the architecture incorporates a 'human-in-the-loop' confirmation mechanism and the Gemini API to facilitate context-aware, bidirectional dialogue. Deployed via a low-latency Streamlit interface, the system provides simultaneous translation into two regional languages (Hindi and Kannada) with synchronised audio output, establishing a highly inclusive and data-driven solution for daily interactions..*

Keywords*:* Sign Language Recognition, YOLOv8, Streamlit, Gesture Detection, Multilingual Translation, Text-to-Speech (TTS), Gemini API, Assistive Technology, Accessibility

I. INTRODUCTION

Communication is fundamental to human interaction, yet individuals relying on sign language face profound barriers when interacting with the general public. Existing Sign Language Recognition (SLR) technologies often suffer from significant limitations, including slow processing speeds, inability to support continuous sentence formation, and poor reliability in dynamic, real-world environments. Furthermore, a critical gap remains in the lack of robust multilingual translation and conversational capabilities in current assistive systems, which prevents seamless interaction across different linguistic groups.

This research addresses these challenges through the development of SignSpeak AI, a comprehensive, real-time system that translates finger-spelling gestures into meaningful text and synthesised speech. The primary objective is to establish an accessible, high-performance platform that accurately identifies ISL alphabet signs using the YOLOv8 deep learning model. Unlike traditional static systems, SignSpeak AI incorporates a user confirmation mechanism for constructing coherent sentences and provides simultaneous translation into regional languages—specifically Hindi and Kannada—with synchronised audio synthesis. Additionally, the system integrates the Gemini API to facilitate natural, AI-driven dialogue, ensuring a seamless and inclusive communication experience deployed via a low-latency Streamlit interface.

II. LITERATURE SURVEY

1. YOLO-Based ISL Recognition: Jadhav et al. [1] utilized a YOLO-based CNN for Indian Sign Language (ISL) detection. While effective for isolated characters, their approach lacks the continuous sentence formation and conversational AI integration that SignSpeak AI provides.

2. Real-Time ISL Recognition Using YOLO-NAS: Kumar et al. [2] focused on optimizing hardware efficiency for mobile devices using YOLO-NAS.

DOI: 10.48175/IJARSCT-30564

3. Comparative Analysis of YOLOv5 and YOLOv7: Sannakki and Rajpurohit [3] demonstrated the superiority of YOLO architectures over traditional CNNs for gesture recognition. SignSpeak AI builds upon this foundation but extends the scope from mere recognition to a complete interactive platform involving TTS and NMT.

4. End-to-End ISL Recognition Pipeline: Gupta et al. [4] proposed a recognition pipeline with user confirmation similar to our approach. However, SignSpeak AI significantly innovates by incorporating regional language support (Hindi/Kannada) and Generative AI (Gemini) for context-aware interaction.

5. YOLOv8 for BISINDO Recognition: Pratama et al. [5] applied YOLOv8 to Indonesian Sign Language with basic TTS. SignSpeak AI enhances this methodology by integrating a chatbot interface, thereby converting a one-way translation tool into a two-way conversational agent.

6. YOLOv8-Nano for ASL Detection: Alshareef et al. [6] emphasized high frame-rate performance on edge devices. While SignSpeak AI maintains low latency, its primary contribution lies in user experience and linguistic inclusivity rather than purely hardware-level optimization.

III. OBJECTIVE

The primary objective is to engineer a low-latency, accessible platform that converts static hand gestures into multimodal communication outputs. Specific aims include:

1. Precision Detection: To implement YOLOv8 for the accurate identification of A–Z ISL alphabets from live video streams.
2. Sentence Construction: To develop a user-verification mechanism that allows the seamless compilation of letters into words and sentences.
3. Linguistic Inclusivity: To provide real-time translation and audio synthesis in English, Hindi, and Kannada.
4. Intelligent Interaction: To integrate the Gemini API for enabling context-aware, two-way dialogue between users.

IV. METHODOLOGY

A. System Overview: In the given fig. 1, the system employs a modular design orchestrated by the Streamlit framework to ensure low-latency performance. The pipeline captures video input, processes frames for gesture detection, allows user- validated sentence formation, and routes the text through translation and speech synthesis engines before displaying the output.

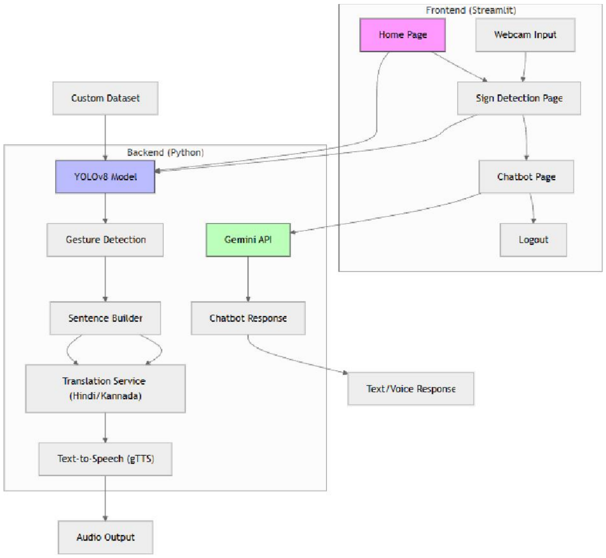

Fig. 1. System Overview of the whole model

B. Data Collection and Preprocessing: A diverse dataset of ISL alphabets was curated using varied camera inputs to account for environmental inconsistencies. To ensure model robustness, the data underwent rigorous preprocessing and augmentation, including random rotations, scaling, and flipping. The images were annotated using LabelImg and exported in standard YOLO format, split into training (70%), validation (15%), and testing (15%) subsets.

C. YOLOv8 Deep Learning Model: The core detection engine utilizes YOLOv8, chosen for its balance of speed and accuracy. The model leverages a CSPDarknet backbone for feature extraction and a Path Aggregation Network (PANet) for feature fusion. Training was conducted via gradient descent to minimize localization and classification loss.

D. Translation, Speech, and Chatbot: The confirmed text is processed by a Neural Machine Translation (NMT) module for conversion into Hindi and Kannada. Simultaneously, a Text-to-Speech (TTS) engine generates natural-sounding audio.

The Gemini API is engaged to handle conversational queries, allowing the system to respond intelligently to the user's signed input.

V. RESULTS

Performance Overview: Evaluation of the SignSpeak AI system indicates high stability and classification accuracy across the ISL alphabet. The model effectively distinguishes between complex hand shapes with minimal latency.

Confusion Matrix:In the fig. 2, the confusion matrix reveals a high density of true positives along the diagonal, signifying accurate class-wise prediction. Misclassifications were rare and largely confined to gestures with high visual similarity (e.g., 'A' and 'E'), suggesting the need for minor dataset refinement.

Training Convergence: Analysis of the training curves demonstrates consistent optimization. Loss metrics (box, classification, and DFL) showed a monotonic decrease, while precision and recall metrics improved steadily.

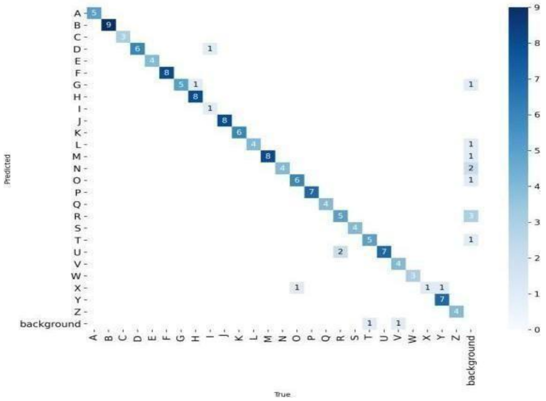

Fig. 2. Confusion matrix Analysis of Alphabet Gesture

VI. ANALYSIS

Model Robustness and Practical Usability: The YOLOv8 model exhibited resilience against environmental variables such as background clutter and lighting shifts. Although performance dipped slightly under extreme occlusion, the system maintained functional reliability for real-time use.

Evaluation Metrics: The system was assessed using standard objective metrics including Mean Average Precision (mAP50 and mAP50-95), Precision, and Recall. Operational efficiency was measured via Frames Per Second (FPS) and latency.

Subjective quality assessments for the Translation and Speech Synthesis modules confirmed that the output was contextually accurate and audibly clear.

VII. LIMITATION

While the YOLOv8 model demonstrates strong performance, certain limitations remain:

- Misclassification rates increase for gestures with highly similar structures.
- Performance may degrade under severe environmental fluctuations (e.g., extreme lighting imbalance or significant motion blur).
- Future work will need to include expanding the dataset with more diverse hand variations, improving preprocessing techniques, and exploring hybrid models combining spatial and temporal features.

VIII. CONCLUSION

The SignSpeak AI project successfully demonstrates a highly effective and inclusive solution for real-time assistive communication. By integrating the YOLOv8 deep learning model, robust multilingual translation, advanced speech synthesis, and the Gemini AI conversational module, the system effectively bridges communication gaps. SignSpeak AI offers a reliable, low-latency, and accessible platform suitable for practical deployment in educational, personal, and accessibility-driven settings. These developments solidify SignSpeak AI's impact, making it a highly adaptive, scalable, and effective solution for diverse real-world communication needs.

IX. ACKNOWLEDGMENT

The authors express their sincere gratitude to their project guide and department faculty for their continuous guidance and support during the completion of this project.

REFERENCES

[1]. Desai K.M., Patil S.S., and Joshi V.A., (2024): Integrating temporal features with YOLO for continuous sign language detection using GRU. IEEE Transactions on Human-Machine Systems, 54(2), 112–125.

[2]. Rodriguez J.M., and Gomez A.B., (2023): Model Compression and Quantization techniques for deploying lightweight YOLO models in assistive communication. Journal of Embedded Systems and Edge Computing, 5(1), 45–55.

[3]. Jadhav S.S., Bairagi V.K., and Pawar S.S. (2023), YOLO Convolutional Neural Network Algorithm for Recognition of Indian Sign Language Gestures, Proc. IEEE Int. Conf. Electron. Syst. Signal Process. Intell. Technol. (ICESSIT), Nagpur, India, 1–6.

[4]. Kumar A., Singh R., and Gupta P. (2024), Indian Sign Language Recognition in Real Time using YOLO NAS, Proc. IEEE Int. Conf. Adv. Computer. Commun. Paradigms (ICACCP), Gangtok, India, 1–6

[5]. Zhang, Z. (2021). Advanced gesture recognition systems using deep learning. In L. Wang & A. B. Smith (Eds.), Computer vision in healthcare (pp. 88-102). Singapore: Springer Nature.

[6]. M. A. P. Taylor (2020). Computer vision for accessibility and assistive technologies. In J. Doe (Ed.), AI for social good (pp. 112-145). Berlin: Springer

DOI: 10.48175/IJARSCT-30564