International Journal of
Human Computations and Intelligence
Vol. 05, Issue. 01, 2026

RESEARCH ARTICLE OPEN ACCESS

Real–Time Sign Language Recognition and Multilingual Speech Output Based on Machine Learning

Lavanya N L* . H R Sujay . Akash M . Darshan S . Akash G A

Department of Computer Science and Engineering,
East West College of Engineering,
Visvesvaraya Technological University, Bengaluru, Karnataka 560064.

DOI: **10.5281/zenodo.17598489**
Received: 18 October 2025 / Revised: 11 November 2025 / Accepted: 13 November 2025
*Corresponding Author: *lavanyaanand96@gmail.com*

Abstract – For people with hearing and speech disabilities, sign language is an essential means of communication. Yet, a communication gap between signers and non-signers remains large because of limited public knowledge. To overcome this limitation, this work proposes a machine learning–based real-time sign language recognition and translation system. The system captures hand movements using a standard webcam and uses the Mediapipe framework to recognize accurate hand landmarks. These landmarks are subsequently categorized using independently trained Random Forest Classifier models for Indian Sign Language (ISL) and American Sign Language (ASL). The identified gestures are translated into text and then audible speech utilizing the pyttsx3 library, and the Google Translate API provides multilingual translation for cross-linguistic communication. Experimental results show that the system proposed performs accurate real-time recognition performance through regular computing hardware alone.

Index Terms – Sign Language Recognition, Mediapipe, OpenCV, Machine Learning, Random Forest, pyttsx3, Google Translate API, ASL, ISL, Human–Computer Interaction.

I. INTRODUCTION

Communication is the basis of social interaction, knowledge sharing, and integration within communities. But in the case of people with speech and hearing disabilities, communication may prove to be challenging within communities where sign language is not recognized. While sign language is expressive and effective[1][2], it is still confined to those who have been formally trained in it [3]. This creates a communication barrier for much of the population and underlines the importance of assistive technologies that can automatically interpret sign gestures. With recent technological

MILESTONE RESEARCH.IN OPEN ACCESS

ISSN (Online): 2583-5696
Int. Jr. of Hum Comp. & Int.

breakthroughs in machine learning, artificial intelligence, and computer vision, it is now possible to develop low-cost real-time sign language recognition systems. Thanks to modern frameworks such as Mediapipe and OpenCV, it is possible to detect and track hand landmarks from a regular webcam without the necessity of pricey sensors or special equipment [4].

The focus of this research is to develop an integrated system that is capable of identifying sign language gestures, constructing meaningful words, translating them into speech, and enabling multilingual translation. Through the use of gesture recognition, natural language processing, and text-to-speech synthesis, the system provides smooth interaction among signers and non-signers [5]. This concept is in favor of accessibility and inclusivity in schools, social, and healthcare settings.

A. Background

Previously, sign language interpretation was reliant on human interpreters or body-mounted technology such as glove-based devices. While these methods worked, they were hindered by being costly, calibration-intensive, and mobility-constrained [6]. Advances in computer vision have significantly transformed this technology with the potential of gesture tracking using camera-based techniques. Google's Mediapipe library provides an incredibly efficient hand-tracking system that has the capability to track 21 important hand landmarks in real time using deep learning models [1].

Previous research has employed deep learning methods like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models in gesture classification [7]. These tend to demand intensive computational resources and large amounts of data. RF models have been competitive for real-time and light applications while being less computationally expensive and being rapid to train [8]. With the incorporation of Mediapipe, OpenCV, and Random Forest classification, this study achieves a balance between accuracy and efficiency. Further, with the incorporation of text-to-speech synthesis (pyttsx3) and translation APIs (Google Translate), inclusion and accessibility are enhanced even more, enabling two-way communication between hearing and hearing-impaired [9].

B. Problem Statement

Despite notable progress in sign recognition technologies, a truly comprehensive and real-time communication system that supports multiple sign languages remains elusive. Existing approaches often suffer from one or more of the following limitations:

- Inability to recognize gestures from both ASL and ISL within a single adaptable framework.
- Inconsistent performance due to background noise, camera variations, or lighting changes.
- Lack of stabilization and word formation logic, resulting in inaccurate or repetitive predictions.
- Absence of integrated speech and translation modules for natural communication.

This project addresses these challenges by developing a system that not only recognizes gestures but also forms complete words, converts them into audible speech, and translates them into multiple languages—all in real time using only a standard webcam.

C. Objectives

- Create an Intelligent Recognition System:Design a real-time sign language translator and recognition system that can accurately identify gestures from both American Sign Language (ASL) and Indian Sign Language (ISL) using Mediapipe and OpenCV for accurate hand landmark detection.
- Implement the integration of Machine Learning and Speech Modules:Train isolated Random Forest Classifier models for ASL and ISL, apply word formation with stabilization logic, and synthesize recognized text to natural speech using the pyttsx3 library.
- Facilitate Multilingual and Interactive Communication:Use the Google Translate API for multilingual translation and design an easy-to-use GUI that shows live video, detected gestures, translated text, and speech output to make it accessible and interactive for users.

II. SYSTEM DESIGN

The proposed system is composed of five major modules that work together to enable real-time recognition, translation, and vocalization of sign language gestures. These modules include Data Acquisition and Preprocessing, Feature Extraction, Model Training, Gesture Recognition with Word Formation, and finally, Speech and Translation. The design ensures modularity, allowing individual components to be updated or improved independently while maintaining the overall functionality and performance of the system.

A. Data Acquisition and Preprocessing

The data acquisition module is responsible for capturing live video input through a webcam using the OpenCV library. Each frame of the video feed is analyzed using the Mediapipe framework, which detects and tracks 21 hand landmarks, providing accurate (x, y, z) coordinate values for each key point. These coordinates represent the position and orientation of the hand in real time.For model training, the system utilizes two primary sources — the Mediapipe Processed ASL Dataset from Kaggle and a custom Indian Sign Language (ISL) dataset created through controlled data collection. During preprocessing, the extracted landmark data is normalized and organized into structured CSV files to ensure uniformity, reliability, and smooth integration with the learning model.

B. Feature Extraction

In the feature extraction phase, the 21 detected hand landmarks are processed to generate numerical feature vectors. Each landmark contributes three coordinates —x, y, and z — resulting in a 63-dimensional feature vector for every captured hand. These vectors effectively capture the geometric structure and spatial configuration of gestures, forming a unique signature for each sign.This standardized numerical representation enables the model to distinguish between different signs efficiently, even when performed by users with varying hand sizes, gestures, or under inconsistent lighting and background conditions.

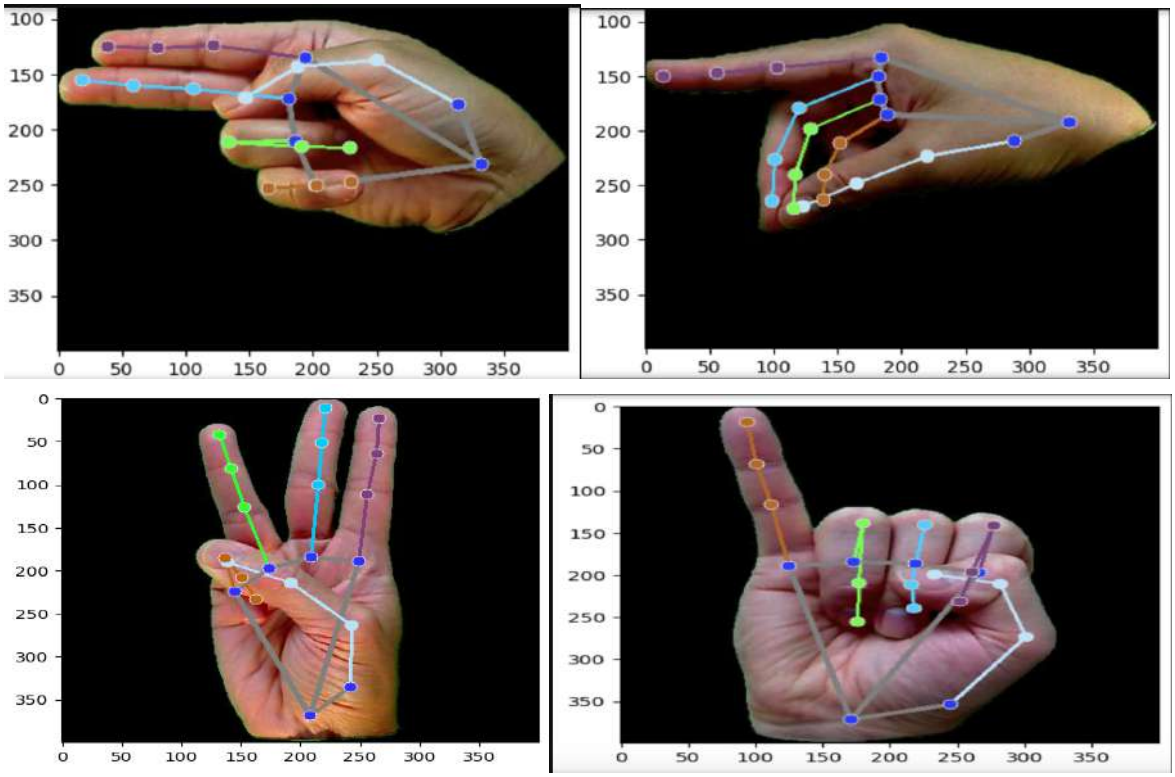

Fig. 1: Recognition of hand gesture and display of landmarks

C. Model Training

The model training phase involves the development of two independent classification models one for American Sign Language (ASL) and another for Indian Sign Language (ISL). The Random Forest Classifier (RFC), a robust and interpretable machine learning algorithm, is employed due to its ability to handle high-dimensional data and resistance to overfitting through ensemble learning. Each model is trained on its respective dataset, where multiple decision trees collectively vote on the final prediction, ensuring improved accuracy and stability. The models are evaluated using accuracy scores and confusion matrix analysis to assess performance and identify areas for refinement. Once trained, the models are serialized and stored for integration into the real-time recognition pipeline.

D. Working Principle of Random Forest

The Random Forest Classifier operates as an ensemble learning algorithm, combining multiple decision trees to improve prediction accuracy and reduce overfitting. It functions by creating a collection of independent decision trees, each trained on a random subset of the dataset and features. When a new input is provided, every tree generates its own classification output, and the final prediction is determined through a majority voting mechanism — the class that receives the most votes is chosen as the output.This approach follows the principle of the *"wisdom of the crowd"*, where the aggregation of multiple weak learners results in a strong overall model.

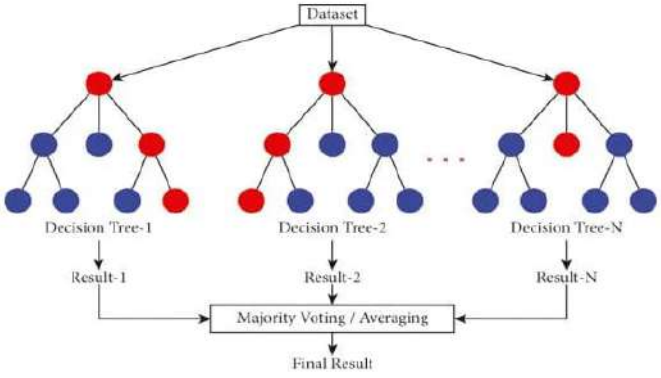

Fig. 2: Random Forest Classifier

In this paper, the Random Forest algorithm was chosen for its robust performance with 63-dimensional landmark data, resilience to noise and variation, and efficiency in real-time inference all of which are essential for accurate sign recognition across diverse users and conditions.

E. Gesture Recognition and Word Formation

During real-time operation, frames captured from the webcam are processed to extract landmark features, which are then passed to the trained Random Forest model for classification. To enhance reliability, a time-based stabilization mechanism introduces a controlled delay between consecutive predictions, reducing false positives and repetitive detections. Recognized letters are sequentially combined to form complete words, bridging the gap between static alphabet gestures and dynamic sentence formation. This ensures a smooth and natural interpretation of user input, making communication more fluid and contextually meaningful in real-world scenarios.

F. Speech and Translation

The final module, Speech and Translation, converts recognized text into both audible speech and translated text output. The pyttsx3 library is employed to generate real-time voice synthesis, enabling instant vocalization of recognized words. Simultaneously, the Google Translate API provides multilingual translation capabilities, allowing the recognized text to be translated into multiple target languages for enhanced accessibility. Together, these functionalities enable cross-linguistic communication and make the system an inclusive tool for bridging the gap between sign language users and non-signers. By transforming gestures into meaningful speech and text, the system completes the gesture-to-speech translation cycle effectively and interactively.

III. RESULTS AND DISCUSSIONS

The developed Sign Language Recognition and Translation System was extensively tested to assess its performance, accuracy, robustness, and usability across both American Sign Language (ASL) and Indian Sign Language (ISL) models. The testing phase focused on evaluating how effectively the system could recognize gestures in real time, form words accurately, and generate both speech and translated text outputs. The results confirmed that the integration of Mediapipe, OpenCV, and the Random Forest Classifier yielded a reliable and responsive solution capable of functioning efficiently on standard computing hardware.

Table. I: Evaluation Parameters

Evaluation parameter	Description / Purpose	Measured Result
Accuracy (%)	Percentage of correctly recognized gestures out of total predictions.	ASL: 93.1% ISL: 88.7%
Precision (%)	Ratio of correctly predicted positive gestures to total predicted positive gestures.	ASL: 92.5% ISL: 87.9%
Recall (%)	Measures how well the model detects all relevant gestures	ASL: 93.8% ISL: 89.4%
F1-Score	Harmonic mean of precision and recall, showing overall classification balance.	ASL: 0.93 ISL: 0.88
Average Latency (sec/frame)	Time required to process a single video frame and display output.	0.05 – 0.08 sec
Word Formation Accuracy (%)	Accuracy in combining detected letters into correct words.	90.4%
Translation Accuracy (%)	Correctness of translated text from Google Translate API.	97.2%

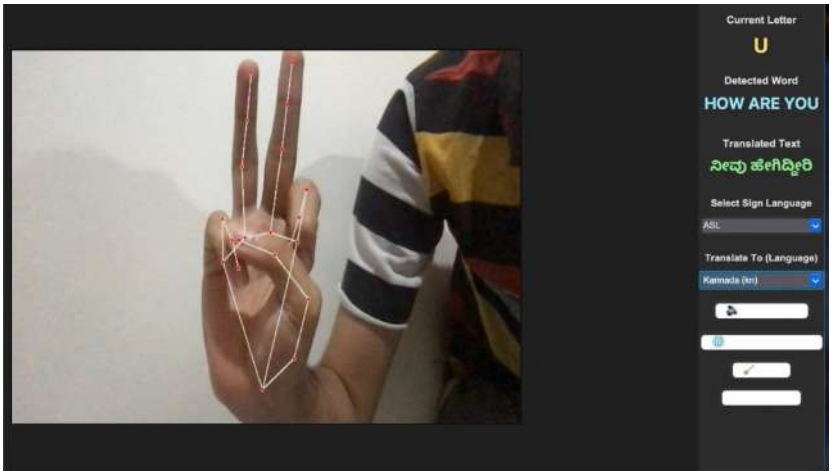

Fig. 3: Multilingual sign-to-text translation in action-Telugu and Japanese examples.

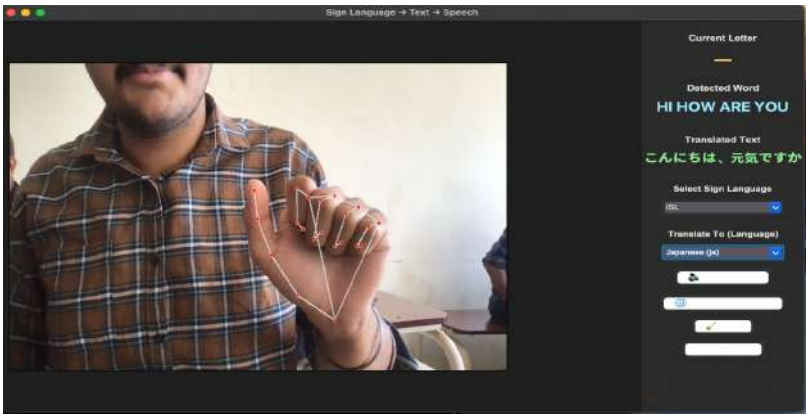

Fig. 4: Multilingual sign-to-text translation in action-Telugu and Japanese examples.

During testing, the models achieved a recognition accuracy ranging between 88% and 93% under standard lighting conditions and against neutral backgrounds. This performance demonstrates the effectiveness of the Random Forest Classifier in distinguishing between subtle hand movements and similar gesture patterns using the 21-hand landmark features extracted by Mediapipe. The models

were further tested under different environmental conditions, including variations in illumination, camera angle, and background noise. Despite these changes, the system maintained consistent performance, with only minor fluctuations in accuracy. This confirms the robustness and generalization capability of the trained models, making them adaptable to real-world scenarios where conditions are rarely uniform.

A key factor contributing to the system's effectiveness is its word formation and stabilization logic, which introduces a short delay between consecutive predictions. This delay prevents rapid fluctuations or repeated alphabet detections, ensuring that transitions between letters appear smooth and natural. As a result, the system produces coherent and readable word outputs, significantly improving user experience and communication clarity. This mechanism effectively bridges the gap between static gesture recognition and continuous, meaningful word generation — a crucial feature for real-time sign interpretation. The real-time prediction pipeline, built with OpenCV for video streaming and Mediapipe for landmark tracking, consistently delivered smooth and stable frame rates on standard systems without the need for GPU acceleration. This confirms the system's computational efficiency and accessibility, making it suitable for deployment even on mid-range laptops or personal computers. The combination of optimized preprocessing, efficient model inference, and controlled frame handling ensures minimal latency, enabling users to communicate without noticeable delays during interaction.

The integration of the pyttsx3 text-to-speech engine added an essential auditory dimension to the system. Each recognized word was instantly converted into clear and natural speech, allowing real-time vocal feedback. Since pyttsx3 operates offline, this feature remains functional even without internet access an important advantage for accessibility and portability. Additionally, the Google Translate API enriched the system's capabilities by providing instant multilingual translation. Recognized words were translated into the user's preferred language, broadening the system's reach and enabling communication across linguistic boundaries. This dual functionality of speech synthesis and translation makes the system not only intelligent but also deeply inclusive. The results demonstrate that the fusion of computer vision, machine learning, and natural language processing technologies creates a highly inclusive and user-friendly communication platform. The system successfully bridges the gap between hearing-impaired users and non-signers, transforming hand gestures into spoken and translated words with impressive accuracy and responsiveness.

Its ability to handle both ASL and ISL gestures, maintain stable performance across varied conditions, and generate multilingual speech output highlights its real-world practicality. The outcomes affirm that this integrated approach not only promotes technological innovation but also fosters social inclusion and accessibility, empowering individuals with speech and hearing impairments to communicate more freely and effectively.

IV. CONCLUSION

The Sign Language Recognition and Translation System that was created uses computer vision, machine learning, and natural language processing to let people who use sign language and people who don't use it talk to each other in real time. It makes sure that gestures are correctly interpreted in

all situations by using Mediapipe and OpenCV for precise gesture detection and a Random Forest Classifier for strong recognition. The system turns gestures into spoken and written outputs right away using pyttsx3 for speech and Google Translate for translation into many languages. It works well in real time without needing high-end hardware, allowing for smooth word formation and very little delay. Deep learning integration (CNN, LSTM), bigger datasets, and deployment on mobile and web are some of the things that will be better in the future. The system makes communication easy, smart, and seamless for everyone, including people who have trouble hearing or speaking.

REFERENCES

1. Gnanapriya, S., & Rahimunnisa, K. (2023). A Hybrid Deep Learning Model for Real Time Hand Gestures Recognition. *Intelligent Automation & Soft Computing, 36(1).*
2. Abdallah, M. S., Samaan, G. H., Wadie, A. R., Makhmudov, F., & Cho, Y. I. (2022). Light-weight deep learning techniques with advanced processing for real-time hand gesture recognition. *Sensors, 23(1), 2.*
3. Yaseen, Kwon, O. J., Kim, J., Jamil, S., Lee, J., & Ullah, F. (2024). Next-gen dynamic hand gesture recognition: Mediapipe, inception-v3 and lstm-based enhanced deep learning model. *Electronics, 13(16), 3233.*
4. 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.*
5. Zhang, Y., Yuan, B., Yang, Z., Li, Z., & Liu, X. (2023). Wi-nn: Human gesture recognition system based on weighted knn. *Applied Sciences, 13(6), 3743.*
6. Lavanya, N. L., Bhat, A., Bhanuranjan, S. B., & Narayan, K. L. (2023). Enhancing the Capabilities of Remotely Piloted Aerial Systems Through Object Detection, Face Tracking, Digital Mapping and Gesture Control. *International Journal of Human Computations & Intelligence, 2(3), 147-158.*
7. Garg, M., Ghosh, D., & Pradhan, P. M. (2023). Multiscaled multi-head attention-based video transformer network for hand gesture recognition. *IEEE Signal Processing Letters, 30, 80-84.*
8. Lavanya, N. L., Savanvur, A. K. V., Shrivatsa, R. S., & Shetty, U. K. (2024). LANE MORPH: Machine Learning Powered Divider For Traffic Volume Adaptation. *International Journal of Human Computations & Intelligence, 3(6), 378-385.*
9. Slama, R., Rabah, W., & Wannous, H. (2025). Online hand gesture recognition using Continual Graph Transformers. *arXiv preprint arXiv:2502.14939.*
10. Kale, H., Aswar, K., Yadav, D. Y. M. K., & Mali, D. Y. (2024). Attendance marking using face detection. *International Journal of Advanced Research in Science, Communication and Technology, 417424.*
11. Ahmed, S. T., Basha, S. M., Arumugam, S. R., & Kodabagi, M. M. (2021). Pattern Recognition: An Introduction. MileStone Research Publications.
12. Kumar, S. S., Ahmed, S. T., Sandeep, S., Madheswaran, M., & Basha, S. M. (2022). Unstructured Oncological Image Cluster Identification Using Improved Unsupervised Clustering Techniques. Computers, Materials & Continua, 72(1).