iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79590

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Real-Time Sign-to-Text Translation System Using Multi-Modal Feature Fusion and Semantic Correction

Rathigha S[1], Pulikanti Rajith Teja[2], Kishore G[3], Dr. Ezhilarasan M[4]
[1, 2, 3]*Student,* [4]*Professor, Department of Information Technology, Puducherry Technological University, Puducherry, 605014, India*

Abstract: ***Communication barriers persist for individuals with hearing and speech impairments because sign language, while expressive, is not widely understood outside the Deaf community. To address this, this paper presents a Real-Time Sign-to-Text Translation System designed to bridge the communication gap for individuals with hearing and speech impairments. Traditional vision-based systems often struggle with continuous gesture recognition, poor accuracy under varying conditions, and a lack of semantic understanding. To overcome these limitations, we propose a three-module architecture. The first module utilizes MediaPipe Holistic for extracting 3D spatial landmarks from the hands, face, and body pose. The second module employs a Long Short-Term Memory (LSTM) network to process these temporal sequences, effectively capturing dynamic motion patterns and stabilizing predictions with confidence gating. The final module integrates a Transformer-based Natural Language Processing (NLP) model alongside deterministic fallback templates to perform semantic correction, converting raw gloss sequences into grammatically coherent English sentences. Experimental results on the LSA64 dataset demonstrate a validation accuracy of 97.2%, with the system sustaining real-time processing capabilities on CPU hardware. The integrated web application delivers low-latency, end-to-end translation, making it a viable assistive technology for inclusive communication in education, healthcare, and public services.***

Keywords: Sign Language Recognition, Multi-Modal Feature Fusion, Semantic Correction, Long Short-Term Memory (LSTM), MediaPipe Holistic, Natural Language Processing.

I. INTRODUCTION

A. Background

Communication barriers persist for individuals with hearing and speech impairments because sign language, while expressive, is not widely understood outside the Deaf community. Vision-based systems have attempted to bridge this gap, but most focus on static hand shapes or isolated gestures. These traditional approaches fail to capture the temporal motion, facial expressions, and context that constitute natural sign language conversations. Furthermore, many prototypes stop at simple classification, leaving users without a real-time, conversational interface.

B. Problem Statement

Despite progress in computer vision and machine learning technologies, existing sign language recognition systems face significant limitations. A major challenge is limited real-time sign recognition; traditional models struggle to interpret continuous sequences of dynamic movements. Another critical issue is insufficient accuracy in recognizing complex gestures. Sign language includes facial expressions, body posture, and subtle motion variations, and existing systems often fail to capture these intricate details under varying lighting conditions or fast hand movements. Additionally, most current systems lack semantic understanding. They perform direct mapping of gestures to predefined labels without considering grammatical structure, leading to incorrect sentences when translating syntax-heavy languages like American Sign Language (ASL).

C. Objective

This project, “Real-Time Sign-to-Text Translation using Multi-Modal Feature Fusion and Semantic Correction,” addresses existing gaps by introducing an end-to-end multi-modal architecture. The primary objective is to provide real-time, accurate, and context-aware translation of sign language into text.

©IJRASET: All Rights are Reserved

By combining full-body cue extraction, temporal motion sequence recognition, and transformer-guided semantic correction, the system aims to output readable English sentences in real time, moving the experience closer to an inclusive conversational aid.

II. LITERATURE REVIEW

In recent years, spatial feature extraction for sign language recognition has been significantly improved using deep learning and multimodal approaches. Traditional methods rely on convolutional neural networks and attention-based architectures to extract visual features such as hand shapes and facial expressions from raw image data [4]. These approaches achieve high recognition accuracy but involve high computational complexity and are not suitable for real-time applications. To enhance performance, multimodal frameworks combining RGB video features and heatmap embeddings have been introduced using advanced architectures such as Conformer-based models [8]. These methods effectively capture both spatial and contextual information for continuous sign language recognition. However, they rely on high-dimensional video data and computationally intensive processing, making them inefficient for lightweight and scalable systems. Alternatively, landmark-based approaches using MediaPipe Holistic have gained attention for efficient feature extraction [14]. These methods convert raw video frames into structured skeletal representations by extracting pose, hand, and facial landmarks. This significantly reduces data dimensionality while preserving essential gesture information. However, existing implementations still focus mainly on classification tasks and lack emphasis on modular, scalable feature extraction pipelines. Overall, the literature highlights a trade-off between accuracy and computational efficiency. While deep learning and multimodal approaches improve recognition performance, they are not suitable for real-time deployment. This justifies the need for a lightweight, landmark-based feature extraction approach, as adopted in our system, to enable efficient and robust real-time gesture recognition.

Temporal sequence learning plays a crucial role in recognizing dynamic sign language gestures by capturing motion patterns over time. Graph-based approaches such as multi-scale spatial-temporal graph convolutional networks (MST-GCN) have been widely used to model skeletal sequences and capture both short- and long-range dependencies [18]. While these methods achieve strong performance in action recognition, they involve high computational complexity and are not suitable for real-time deployment.To improve temporal modeling efficiency, recurrent neural network-based approaches such as Long Short-Term Memory (LSTM) networks have been introduced for dynamic gesture recognition [17]. These models utilize sequential landmark data, often extracted using MediaPipe, and incorporate attention mechanisms to enhance temporal feature learning. Although optimized for edge devices, such systems still suffer from higher inference latency and limited scalability for continuous recognition tasks.Further advancements include skeleton-based deep learning frameworks that combine architectures such as CNN-LSTM, Bi-LSTM, and GRU for sequence-based sign language recognition [9]. These approaches improve classification accuracy by leveraging temporal dependencies in landmark data. However, they primarily focus on isolated gesture recognition and are not well-suited for real-time continuous sign language translation.Overall, the literature highlights the importance of efficient temporal modeling techniques. While graph-based and hybrid deep learning models improve accuracy, they lack real-time efficiency and scalability. This justifies the use of a lightweight LSTM-based sequence learning approach in our system for effective and real-time gesture classification.

Recent advancements in sign language translation and natural language processing have leveraged transformer-based architectures for generating meaningful text outputs. Large-scale models such as gloss-free sign language translation frameworks utilize Vision Transformers and GPT-based language models to directly convert sign inputs into natural language sentences [12]. These approaches eliminate the dependency on intermediate gloss representations and achieve high translation accuracy. However, they require large pretrained models and significant computational resources, making them unsuitable for lightweight real-time applications.To improve sentence quality, transformer-based multilingual models have been proposed for grammatical error correction by treating the task as a machine translation problem [2]. These models use pretrained language models to refine sentence structure and improve linguistic correctness. While effective for text correction, such approaches focus only on textual inputs and do not integrate multimodal data or real-time sign language processing pipelines.Additionally, transformer-based models such as BART and T5 have been widely applied for text generation and summarization tasks using encoder–decoder attention mechanisms [10]. These models demonstrate strong performance in generating coherent and context-aware text outputs. However, they are primarily designed for standalone text processing tasks and lack integration with real-time systems or sign language translation frameworks.Overall, the literature highlights that while transformer-based models provide powerful capabilities for language generation and correction, they are often computationally intensive and lack real-time system integration. This justifies the use of a lightweight transformer-based NLP module in our system, combined with web-based deployment, to enable efficient semantic correction and real-time user interaction.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

III. PROPOSED SYSTEM

A. System Architecture

The proposed system relies on a continuous live RGB video stream, processed in real time without storing raw footage. The architecture is divided into three distinct modules. First, the video stream is fed into the spatial feature extraction layer, which captures detailed skeletal landmarks. These landmarks are normalized into temporal sequences and passed to an LSTM-based gesture classification module designed to capture temporal dependencies and motion patterns. The stabilized sequence of recognized signs (glosses) is then processed by a Transformer-based NLP module for semantic correction. Finally, the system is integrated into a Flask-based web application via WebSocket for real-time frontend display. (See Fig. 1 for High-Level Architecture).

B. Spatial Feature Extraction and Landmark Tracking

Module I acts as the primary perception layer. Each captured frame is processed using the MediaPipe Holistic framework, which detects 33 pose landmarks, 468 facial landmarks, and 21 landmarks for each hand. Instead of processing raw pixel values, the system normalizes these spatial coordinates relative to frame dimensions to handle variations in camera distance and user positioning. The coordinates are concatenated into a fixed-dimension feature vector of size 411, representing the complete spatial configuration of the signer for each frame.

C. Temporal Sequence Learning and Gesture Classification

Module II interprets the spatio-temporal landmark data. The incoming 411-dimensional feature vectors are segmented using a fixed-length sliding window (typically 30 frames) stored in a First-In-First-Out (FIFO) buffer. Before classification, motion and visibility gating ensures hand landmarks are visible and motion exceeds a predefined threshold. Validated sequences are fed into a two-layer LSTM network (hidden size 128, with dropout). The final hidden state is passed through a fully connected layer and a softmax classifier to generate probability scores for predefined sign glosses. A confidence gating mechanism combined with a voting buffer ensures that a gloss is finalized only when it consistently dominates recent predictions.

D. Semantic Correction and Web Integration

Module III bridges machine recognition and human understanding. Raw glosses are normalized to collapse consecutive duplicates and preserve multi-word tokens. The sequence is input into a lightweight T5-small sequence-to-sequence model to generate fluent English sentences. If the transformer output is empty or unclear, deterministic fallback templates guarantee a coherent result based on known vocabulary mappings. The final sentence, alongside gloss history and skeletal overlays, is delivered in real time to the user interface via a Flask backend and WebSocket communication.

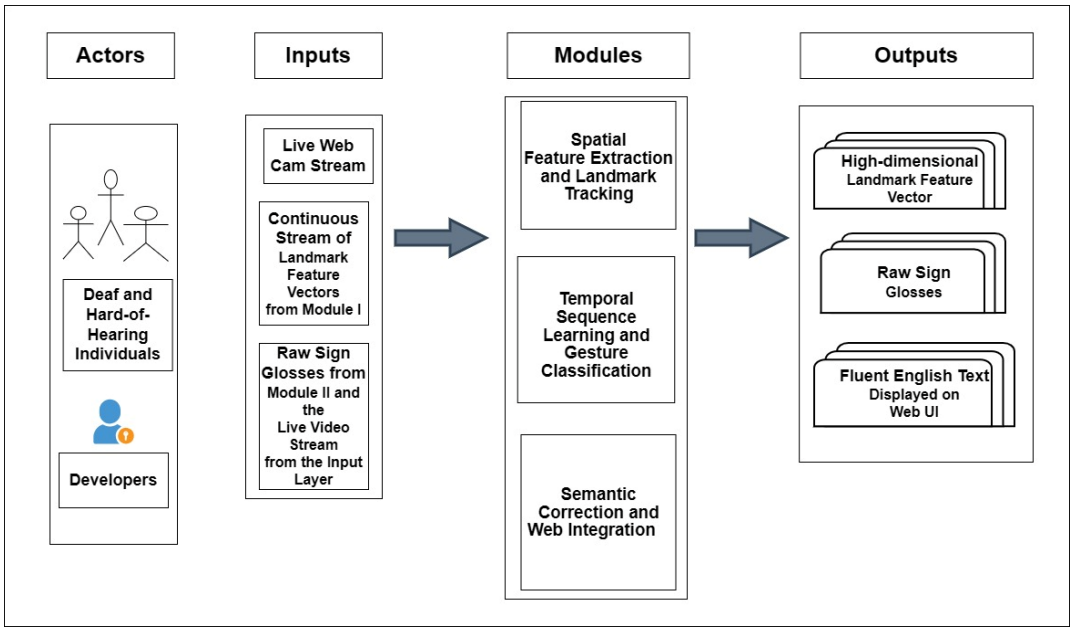


Fig.1 High-Level Architecture of the Proposed System

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

IV. RESULTS AND DISCUSSION

A. *Comparative Analysis*

Traditional sign-language recognition systems fall into two broad categories:

1) static alphabet detectors based on CNNs/HOG features, and

2) frame-based classifiers that lack temporal modeling.

Such systems perform adequately on carefully posed, static hand shapes but fail when confronted with continuous signing, facial expressions, or body posture changes. They also stop at gloss labels, leaving the user to interpret raw outputs.

The proposed system advances beyond these limitations in three ways:

- Multi-modal input – Instead of relying solely on hand crops, the pipeline fuses pose, facial, and hand landmarks, capturing non-manual signals essential for natural sign language.
- Temporal modeling – A sliding window plus LSTM captures motion trajectories, enabling recognition of dynamic signs that static models miss. Confidence thresholds and voting smooth the output, yielding cleaner gloss sequences.
- Semantic correction – The transformer + template layer translates gloss streams into grammatical sentences, something rarely addressed by earlier models. This makes the output immediately usable, especially for hearing users unfamiliar with glossary notation.

Feature	Prior Systems (YOLOv8, CNN+HOG)	Proposed System
Input Modalities	Hand images only	3D pose + face + hands
Temporal Understanding	Limited or none	LSTM with motion gating
Output	Gloss labels/alphabet	Gloss + English sentence
Real-time Deployment	Often GPU-bound	CPU-friendly (MediaPipe + LSTM + T5)

Table1:Comparative Analysis with Existing Systems

B. *Evaluation Metrics*

To demonstrate both recognition quality and real-time usability, we report metrics across three layers of the system.

1) Offline Accuracy (Dataset-Based)

a) LSA64 (64 classes, 3,200 clips)

- Train/val split: 2,560 / 320.
- Best validation accuracy: 97.19 % at epoch 40.
- Full-dataset sweep using cached sequences logs per-class accuracy (50 clips per gloss). Most glosses remain above 95 %; ambiguous ones (e.g., RICE vs. CALL) are flagged for retraining.
- CSV output provides video_path, true_label, pred_label, confidence, and top-k probabilities for forensic analysis.

b) MS-ASL Top-6 (124 train / 26 val)

- Best validation accuracy: 74.36 %.
- Reflects limited data but confirms the same LSTM architecture generalizes to smaller vocabularies.

c) How2Sign subsets Used mainly for sequence-length tuning; qualitative checks show consistent gloss detection for longer sentences.

2) Real-Time Performance Metrics

a) Feature Extraction Latency:

- MediaPipe Holistic processes each 720p frame in ~30 ms on CPU, yielding ~30 FPS landmark streams.

b) LSTM Inference Latency:

- With a 30-frame window, forward pass + softmax takes <50 ms (PyTorch CPU). Because windows overlap, new predictions emerge roughly every frame once the buffer is full.

c) End-to-End Delay:

- From gesture completion to sentence display:Buffer fill (~1 second for 30 frames at 30 FPS) + LSTM inference (<50 ms) + NLP translation (~70–80 ms) = ≈1.1–1.2 seconds perceived delay, acceptable for conversational demo scenarios.

©IJRASET: All Rights are Reserved

d) Throughput:

- The pipeline sustains 25–30 FPS landmark processing and produces 1–2 glosses per second depending on the signer and cooldown settings.

C. Output Quality & Stability Metrics

1) Confidence Statistics:

- Average confidence of emitted glosses >0.85, margin >0.25 thanks to thresholding. Gatedframes (insufficient hand visibility) represent ~15–20 % of total frames, preventing spurious outputs.

2) NLP Reliability:

- Transformer pathway (T5-small) succeeds on >90 % of gloss sequences, producing fluent English in <80 ms.
- Rule-based fallback covers 100 % of MS-ASL top-5 and LSA64 top-64 glosses, ensuring no empty sentences reach the UI.

3) UI Diagnostics:

- Real-time counters display pose/face/hand landmark detection rates (>90 % when signer staysin frame), FPS, and mode ("live" vs. "mock"), helping operators verify health during demos.

D. Graph output Comparative analysis with Existing

This bar chart (fig.2) contrasts our multi-modal LSTM against static-only baselines. YOLOv8 alphabet detectors and CNN+HOG pipelines hover around 90 % and 85 % accuracy, respectively, because they rely on single-frame hand shapes. The proposed approach reaches 97.2 % validation accuracy by fusing pose, face, and hand landmarks with temporal modeling, demonstrating a clear performance gap over existing methods.

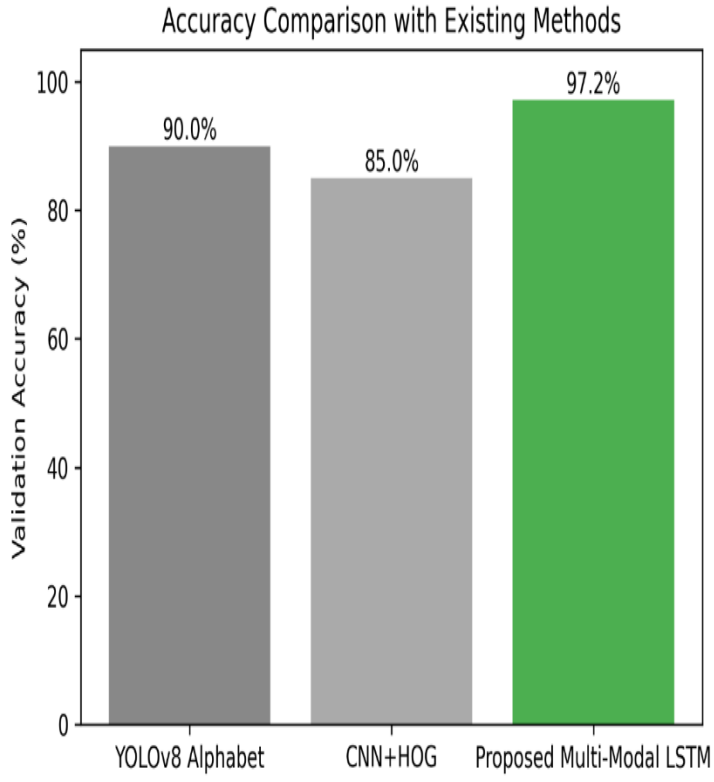


Fig2:Accuracy Comparison

This plot decomposes the end-to-end delay per module. Feature extraction via MediaPipe averages ~30 ms/frame, LSTM inference on a 30-frame window adds ~50 ms, and the NLP layer contributes ~80 ms. Summed together (≈160 ms plus buffer fill time), the pipeline stays within real-time constraints on CPU hardware, validating that higher accuracy does not sacrifice responsiveness.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

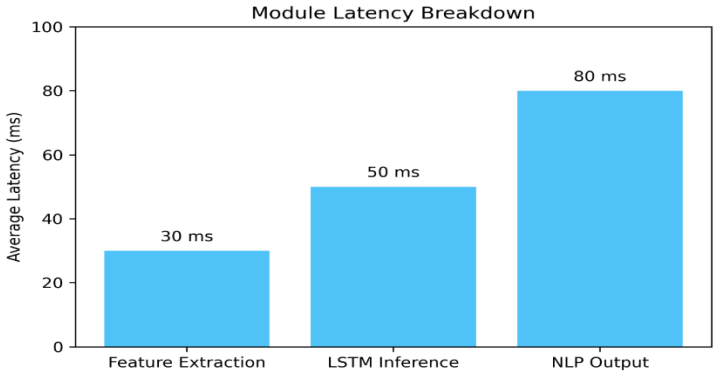


Fig3:Latency Breakdown

This chart plots the accuracy achieved on each of the 64 LSA signs during the cached evaluation run. Even the weakest classes—Red (002), Bright (005), Opaque (001), and Candy (046)—remain above 60 %, while the majority of glosses sit between 90 % and 100 %. The distribution confirms that the model's overall 97.3 % accuracy is not driven by a few easy labels; rather, performance is consistently high across the vocabulary. The handful of lower-scoring classes highlight where targeted data augmentation or signer-specific fine-tuning could further tighten the curve.

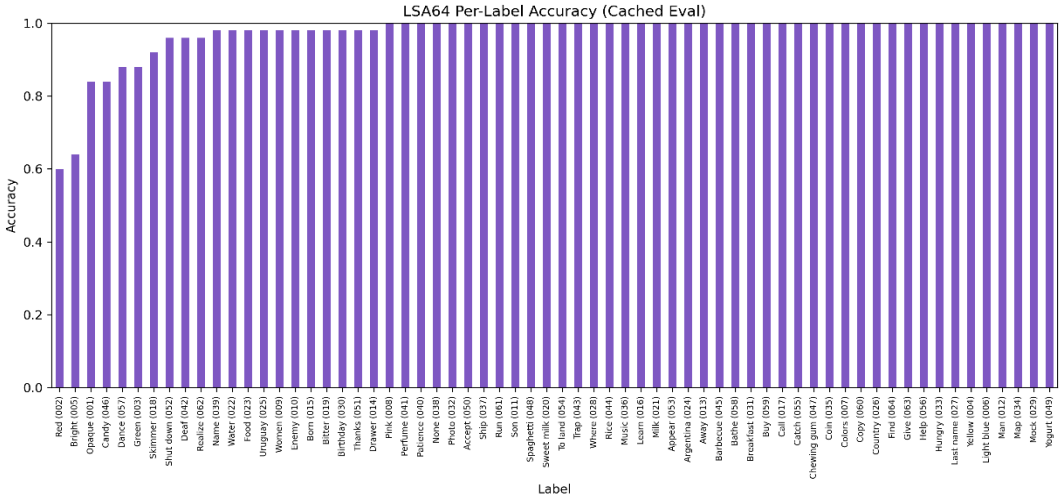


Fig4: LSA64 Per-Label Accuracy

V. CONCLUSION

This project successfully develops a real-time sign-to-text translation platform that effectively bridges the communication gap for individuals relying on sign language. By representing frames as fused 411-dimensional feature vectors, capturing temporal dynamics via an LSTM network, and refining raw predictions through a transformer-based NLP layer, the system provides accurate, grammatically correct English sentences. Achieving a 97.2% validation accuracy on the LSA64 dataset and maintaining robust real-time performance, the system marks a significant step in transforming sign language recognition into a deployable, user-friendly assistive technology.

VI. FUTURE SCOPE

Future enhancements include multi-dataset fusion for automated model selection based on user preference and the integration of advanced temporal models like bidirectional LSTMs or lightweight vision transformers. Furthermore, applying model optimization techniques such as quantization and TFLite/ONNX will enable faster offline execution on mobile and edge devices. Continuous collection of user feedback and analytics will aid in retraining models to consistently improve system accuracy over time.

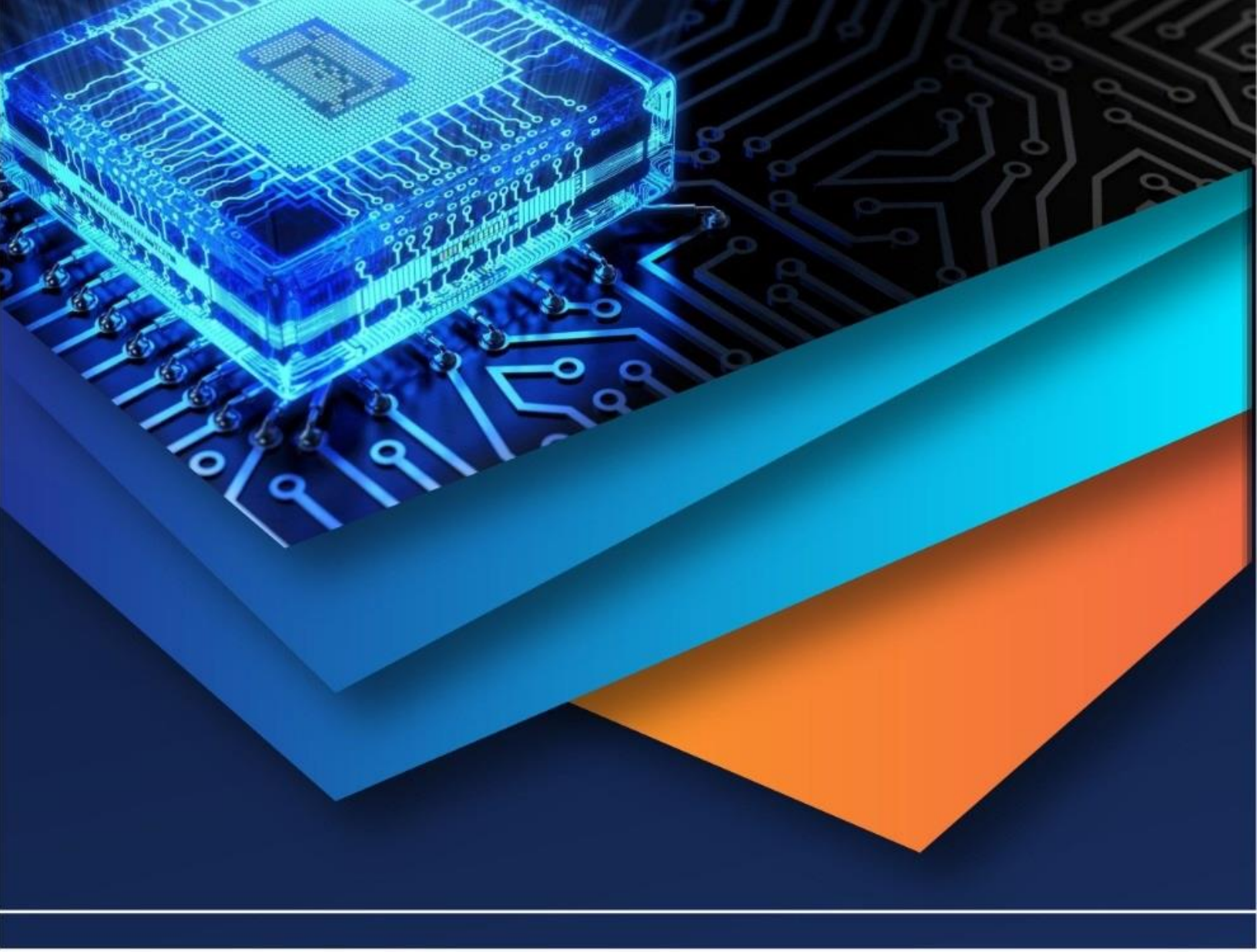
VII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Puducherry Technological University , especially the Department of Information Technology, for providing the necessary infrastructure and support throughout the development of this project. Special thanks to our guide, Dr.M.Ezhilarasan, Professor - Puducherry Technological University for her valuable guidance, constant encouragement, and insightful suggestions that helped shape this research. We would also like to thank our peers and the volunteers who participated in the user testing phase, whose feedback greatly contributed to improving the system. Finally, we appreciate our families and friends for their unwavering support and motivation throughout this journey.

REFERENCES

[1] A. S. M. Miah, M. A. M. Hasan, Y. Tomioka, and J. Shin, “Hand Gesture Recognition for Multi-Culture Sign Language Using Graph and General Deep Learning Network,” IEEE Open Journal of the Computer Society, vol. 5, pp. 144–156, 2024.
[2] A. Luhtaru, E. Korotkova, and M. Fishel, “No Error Left Behind: Multilingual Grammatical Error Correction with Pre-trained Translation Models,” in Proc. EACL, 2024.
[3] B. Alsharif, E. Alalwany, and M. Ilyas, “Transfer learning with YOLOV8 for real-time recognition system of American Sign Language Alphabet,” Franklin Open, vol. 8, pp. 1–11, 2024.
[4] E. Yenisari and S. Yavuz, “Deep Learning-Based Sign Language Recognition Using Efficient Multi-Feature Attention Mechanism,” IEEE Access, vol. 13, pp. 126684–126702, 2025.
[5] M. R. Hassan, K. Nordin, and M. R. Islam, “A review on deep learning techniques for sign language recognition,” IEEE Access, vol. 9, pp. 101789–101810, 2021.
[6] M. Zhang, S. Yang, and M. Zhao, “Deep Learning-Based Standard Sign Language Discrimination,” IEEE Access, vol. 11, pp. 125822–125835, 2023.
[7] M. Al-Qurishi, T. Khalid, and R. Souissi, “Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues,” IEEE Access, vol. 9, pp. 126917–126951, 2021.
[8] N. Aloysius, G. M., and P. Nedungadi, “Optimized Multi-Modal Conformer-Based Framework for Continuous Sign Language Recognition,” IEEE Open Journal of the Computer Society, vol. 6, pp. 739–749, 2025.
[9] P. Antonowicz, D. Kasperek, and M. Podpora, “Sign Language Recognition—Dataset Cleaning for Robust Word Classification in a Landmark-Based Approach,” IEEE Access, vol. 13, pp. 81877–81888, 2025.
[10] R. Rao, S. Sharma, and N. Malik, “Automatic Text Summarization Using Transformer-Based Language Models,” International Journal of System Assurance Engineering and Management, vol. 15, no. 6, pp. 2599–2605, 2024.
[11] R. Varghese and S. M., “YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness,” in Proc. ADICS, Chennai, India, pp. 217–221, 2024.
[12] R. Wong, N. C. Camgoz, and R. Bowden, “Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation,” in Proc. ICLR, 2024.
[13] S. Alyami, H. Luqman, and M. Hammoudeh, “Isolated Arabic Sign Language Recognition Using a Transformer-Based Model and Landmark Keypoints,” ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 1, 2024.
[14] S. K. Anithadevi, S. K. Palanisamy, S. S. Rubini, and S. Shrestha, “MediaPipe-LSTM-Enhanced Framework for Real-Time Dynamic Sign Language Recognition in Inclusive Communication Systems,” Engineering Reports, Wiley Online Library, 2025.
[15] S. Mitra and T. Acharya, “Gesture Recognition: A Survey,” IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 37, no. 3, pp. 311–324, May 2007.
[16] T. Starner and A. Pentland, “Real-Time American Sign Language Recognition from Video Using Hidden Markov Models,” in Proc. International Symposium on Computer Vision, pp. 265–270, 2002.
[17] V. Sharma, A. Sharma, and S. Saini, “Real-time attention-based embedded LSTM for dynamic sign language recognition on edge devices,” Journal of Real-Time Image Processing, vol. 21, article 53, 2024.
[18] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, “Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition,” in Proc. AAAI Conference on Artificial Intelligence, pp. 1113–1122, 2021.
[19] Z. Chen et al., “C2RL: Content and Context Representation Learning for Gloss-Free Sign Language Translation and Retrieval,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 9, pp. 8533–8545, 2025.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)