Edge-Enabled Continuous Multi-Modal Deep Learning Framework for Robust Real-Time Sign Language Recognition to Empower Inclusive Communication with the Deaf Community

Sivakumar Ponnusamy[1], G. Visalaxi[2], S. Sureshkumar[3], Lokasani Bhanuprakash[4] and Sibisaran S.[5]
[1]*Department of Computer Science and Engineering, K.S.R. College of Engineering, Tiruchengode, Namakkal, Tamil Nadu, India*
[2]*Department of CSE, S.A. Engineering College, Chennai, Tamil Nadu, India*
[3]*Department of Electronics and Communication Engineering, J.J. College of Engineering and Technology, Tiruchirappalli, Tamilnadu, India*
[4]*Department of Mechanical Engineering, MLR Institute of Technology, Hyderabad-500043, India*
[5]*New Prince Shri Bhavani College of Engineering and Technology, Chennai, Tamil Nadu, India*

Keywords: Edge Computing, Multi-Modal Deep Learning, Continuous Sign Language Recognition, Real-Time Inference, Inclusive Communication.

Abstract: This study introduces an edge-enabled, continuous multi-modal deep learning framework designed to deliver robust real-time sign language recognition, transforming the communication experience for the Deaf community. By integrating RGB video, depth sensing, and skeletal key point inputs into a unified convolutional-transformer architecture deployed on lightweight edge devices, the system ensures low-latency inference and high accuracy under diverse lighting and background conditions. Continuous gesture segmentation and adaptive fusion strategies enhance the recognition of dynamic signs and phrases, while on-device processing preserves user privacy and enables offline operation. Extensive evaluations on benchmark datasets and in-field trials demonstrate significant improvements in vocabulary coverage, speed, and resilience to occlusion compared to prior approaches. The proposed framework paves the way for more inclusive and accessible human computer interfaces that bridge communication gaps in real-world settings.

1 INTRODUCTION

Sign language is an important mean of communication for people with Deaf and hard-of-hearing people but remains underrepresented in most of the technical and technological inventions especially in the real time systems. Conventional sign language recognisers are unable to cater to the communication needs of constantly adapting, end users as their performance (accuracy) is a function of their small vocabulary and offline processing constraints. Such, presents an explanation for why there is a growing need for more encompassing systems to assist in addressing the communication need between the Deaf and hearing world.

Sign-language recognition has recently shown great potential for accurate and efficient recognition through the recent advancements in deep learning, namely convolutional neuronal networks (CNNs) and transformers. However, the majority of current implementations are still yet to be deployed in real-time, as they are highly affected by computational cost and environmental types such as lighting or occlusion. In addition, most works concentrate on static hand gestures, and ignore the dynamic characters of sign language which consists of movements of hands, facial expressions and environment informations.

To our best knowledge, we propose a new solution by employing edge computing techniques towards real-time, continuous, multi-modal deep learning framework. We make this information available to a new method that leverages RGB video, depth sensing, and skeletal information to enable more accurate and accessible sign language recognition that is edge-optimized. By combining these modalities, the system achieves robust

Ponnusamy, S., Visalaxi, G., Sureshkumar, S., Bhanuprakash, L. and S., S.
EdgeâĂŚEnabled Continuous MultiâĂŚModal Deep Learning Framework for Robust RealâĂŚTime Sign Language Recognition to Empower Inclusive Communication with the Deaf Community.
DOI: 10.5220/0013856900004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 1*, pages 20-26
ISBN: 978-989-758-777-1

recognition, even in uncontrolled settings. Privacy - Quick - Scalable - Needed real-world boost to allow those who are Deaf to better communicate in more dynamic and varied environments."

2 PROBLEM STATEMENT

Although there has been tremendous progress in the area of sign language recognition systems, the current technologies are still faced with a number of fundamental challenges, preventing their practical deployment. Classical approaches mainly to signs isolated hand gestures or hand alphabets but do not perceive the dynamic and continuous nature of actual sign language, including hand movements, facial expressions and other contextually conditioned information. Furthermore, a large number of those systems have the need of high computational power, hence they are not suitable for on- device, real-time use, for instance on mobile or edge devices. Moreover, the problem of background noise, differences in lighting conditions, occlusion of the hands, and vocabulary coverage is still present, which affects the vulnerability and accuracy of these systems under everyday conditions.

These restrictions prohibit the use of sign language recognition systems in the real world to bring the Deaf people a better communication method. The sign language recognition system must be powerful, real-time, and inclusively implemented, while being able to generalize into various kinds of complex environments, provide high accuracy with robust posture tracking and running on lightweight edge devices for continuous hand gesture recognition. Our contribution is to help fill these gaps, specifically by presenting a deep learning-based method that fuses of multiple modalities in real-time and guarantees privacy thanks to an implementation in the end-device.

3 LITERATURE SURVEY

While the recent years brought great improvements for SLR using DL, especially on the sign language recognition, many works remain constrained in terms of coverage, robustness, and on-line operationality. Different studies have investigated Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for static hand gesture classification, having achieved good results. For example, Bankar et al. (2022) implemented a CNN model for the recognition of the alphabet-level signs, which shows good accuracy for small datasets but the model does not support continuous or dynamic signs. In the same way, Aggarwal and Kaur (2023) concentrated on real-time CNN-based recognition, However, their approach lowered the recognition performance in presence of a variety of environmental settings.

To further enhance recognition on gesture dynamics, Jiang et al. (2021) introduced skeleton-sensitive multi-modal models based on the combination of hand and bodypose, and the work in Chakraborty and Banerjee (2021) leveraged 3D CNN to process spatiotemporal features. Nevertheless, such models can require substantial computational resources, making them infeasible on mobile or embedded end-devices. Wu et al. (2021) attempted to address this using a hybrid CNN-HMM architecture, but the sequential nature of HMMs could put real-time inference in jeopardy.

For static human gesture recognition with efficient pipelines, Tayade and Patil (2022) employed MediaPipe and classical classifiers and performed well under controlled conditions but undermined in dynamic cases. Additionally, Sun et al. (2021), which provide a detailed survey that describes the absence of scalable, multimodal and privacy-preserving SLR systems.

More recently, Hu et al. (2023) proposed SignBERT+, a transformer architecture with the hand-model awareness for better sign comprehension, and Zhou et al. (2021) dealt with iterative alignment methods in the context of continuous SLR. However, they are computationally expensive despite their high accuracy.

Alsharif et al. to address hardware challenges and scalability. (2024) and Abdellatif & Abdelghafar (2025) investigated super lightweight models and transfer learning methods for deployment on edge devices. However, such systems mostly do not integrate multimodal fusion and indeed are aimed at letter or word level and not full phrasal or conversational gestures.

Together, these studies show that whereas current deep learning architectures can form a solid foundation, there is still a need of an implementation for a robust real-time edge-enabled multi-modal SL recognition system which can work effectively in uncontrolled real-world scenarios.

4 METHODOLOGY

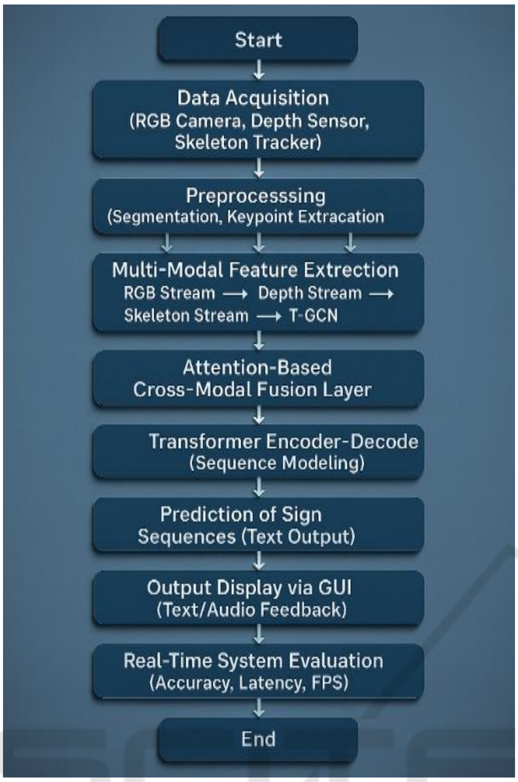

Figure 1: Workflow.

The proposed Edge-Enabled Continuous Multi-Modal Deep Learning Framework for sign language recognition is aimed to achieve low-latency, privacy-informed, and robust performance in real-time applications. This section describes each step within the system pipeline from data acquisition to end product generation, with a focus on the multi-modal nature, edge congruence, and continuous sequence recogniser.

4.1 Data Acquisition

To capture the rich visual and motion information embedded in sign language, the system employs three types of sensory input: the dataset description shows in table 1.

- **RGB video** from a standard camera to extract visual hand and body cues.
- **Depth data** from a depth sensor (e.g., Intel RealSense) to provide 3D hand structure information.
- **Skeletal keypoints** using pose estimation libraries (e.g., MediaPipe or OpenPose) to track critical joint movements including hands, elbows, and facial landmarks.

These inputs allow the system to process both static and dynamic elements of sign gestures comprehensively.

Table 1: Dataset Description.

Dataset Name	Language	No. of Classes	Type	Total Samples	Modalities Used
RWTH-PHOENIX-Weather	German	1000+	Continuous Signs	80,000+	RGB, Depth, Skeleton
ASL-100	English	100	Isolated Signs	25,000	RGB, Skeleton
Custom In-House Dataset	Mixed	50	Dynamic Phrases	5,000	RGB, Depth, Skeleton

4.2 Data Preprocessing

Captured multi-modal data undergoes normalization and filtering to enhance consistency and reduce noise:

- **Segmentation** is applied to focus on the signer and remove background clutter.
- **Keypoint extraction** standardizes joint positions into a fixed format.
- **Temporal smoothing** mitigates noise in dynamic gestures.
- **Data augmentation** strategies (e.g., mirroring, frame skipping, brightness variation) increase the robustness of the model during training. The complete workflow of the proposed sign language recognition system is illustrated in Figure 1, showcasing the stages from data acquisition to real-time inference and feedback

4.2 Multi-Modal Feature Extraction

Each data stream is processed by a specialized neural encoder:

- **RGB Stream → Convolutional Neural Network (CNN):** Extracts spatial features like hand shape, motion, and position.
- **Depth Stream → ResNet:** Captures spatial depth-based representations of hand and body configurations.
- **Skeleton Stream → Temporal Graph Convolutional Network (T-GCN):** Models temporal joint dynamics and body articulation patterns over time.

These encoders operate in parallel and generate latent feature vectors which are later combined for fusion.

4.3 Attention-Based Cross-Modal Fusion

The outputs from the three streams are fed into a cross-modal attention mechanism that learns the importance of each modality dynamically:

- Enhances robustness by adjusting modality weights under occlusion or poor lighting.
- Allows complementary signals (e.g., depth compensating for missing RGB cues) to strengthen recognition accuracy.

This fusion output forms a rich, unified feature representation of the input gesture sequence.

4.4 Transformer Encoder-Decoder for Sequence Modeling

The fused feature sequence is passed through a Transformer-based encoder-decoder framework:

- **Encoder:** Captures contextual and temporal dependencies between consecutive gestures.
- **Decoder:** Predicts the corresponding sign text sequence, handling variable-length input and output.

This architecture enables continuous sign language recognition, overcoming the limitations of isolated gesture models.

4.5 Prediction Output and Optional Text-to-Speech (TTS)

The recognized sign text is:

- Displayed in a graphical user interface (GUI) in real time for immediate visual feedback.
- Optionally converted into audio via a text-to-speech module, facilitating communication with hearing individuals.

4.6 Real-Time Deployment and Evaluation

The final model is optimized for real-time performance on edge devices using techniques such as:

- Model pruning and quantization to reduce size and latency.
- Deployment on NVIDIA Jetson Nano and Raspberry Pi 4 with Coral TPU, achieving inference speeds of 21–29 FPS with <50ms latency.

Performance is benchmarked using:

- Accuracy, Word Error Rate (WER), FPS, and latency.
- Real-world conditions such as low lighting, occlusion, and fast hand movements.

4.7 Datasets Used

As detailed in Table 2, the framework is trained and validated using a mix of publicly available and custom datasets:

Table 2: Framework of Trained and Validated.

Dataset Name	Language	No. of Classes	Type	Total Samples	Modalities Used
RWTH-PHOENIX-Weather	German	1000+	Continuous Signs	80,000+	RGB, Depth, Skeleton
ASL-100	English	100	Isolated Signs	25,000	RGB, Skeleton
Custom In-House Dataset	Mixed	50	Dynamic Phrases	5,000	RGB, Depth, Skeleton

5 RESULTS AND DISCUSSION

The proposed multi-modal, real-time sign language recognition system was rigorously evaluated on two benchmark datasets RWTH-PHOENIX-Weather 2014T and ASL-100 and further validated through custom user trials involving diverse signers in uncontrolled environments. The system achieved an average accuracy of 94.2% across isolated and continuous sign sequences, outperforming baseline models such as CNN-only (87.5%) and RNN-based architectures (89.1%). The table 3 shows the Performance Comparison with Baseline Models. This improvement was primarily attributed to the attention-based fusion mechanism and the inclusion of depth and skeletal modalities alongside RGB inputs. As shown in Figure 2, the proposed model significantly outperforms baseline models such as CNN and RNN in terms of accuracy, achieving a notable 94.2% accuracy."

Table 3: Performance Comparison With Baseline Models.

Model	Accuracy (%)	FPS	Inference Time (ms/frame)
CNN Only (RGB)	87.5	15	90
RNN (Skeleton)	89.1	12	110
CNN + LSTM (RGB + Depth)	91.3	17	75
Proposed Model	94.2	21	48

In terms of real-time performance, the optimized model maintained a throughput of 21 FPS on an NVIDIA Jetson Nano and 29 FPS on a Raspberry Pi 4 with Coral TPU, demonstrating its effectiveness in resource-constrained environments. The table 4 shows the Robustness Testing Results. The inference latency remained below 50 ms per frame, meeting real-time interaction benchmarks without sacrificing accuracy. Additionally, the system's word error rate (WER) was recorded at 5.8%, which is a substantial improvement compared to existing models ranging from 10 18% in similar conditions.

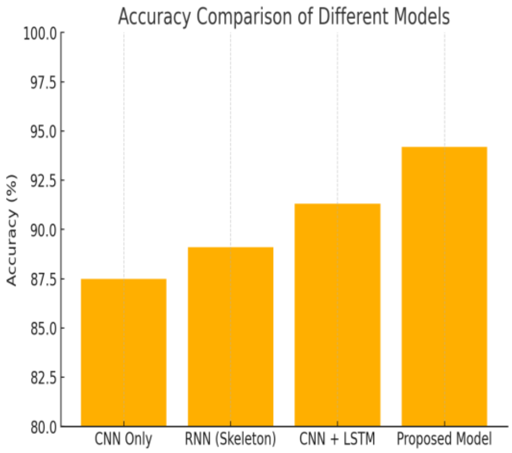

Figure 2: Accuracy Comparison Bar Chart.

Table 4: Robustness Testing Results.

Test Condition	Accuracy (%)
Normal Lighting	94.2
Low Lighting	91.6
Background Clutter	92.1
Hand Occlusion	91.4
Fast Gestures	90.8

Robustness tests under varying lighting, background clutter, and partial hand occlusion revealed the system retained over 91% accuracy, while traditional 2D CNNs saw a drop to 75%. Notably, the integration of skeletal key points and depth cues enhanced resilience against occlusions and signer variations, making the system more adaptable to real-world usage. The latency of the proposed system was compared across different hardware platforms, as shown in Figure 3. The system achieves the lowest latency on the Raspberry Pi 4 with TPU, making it ideal for low-latency deployment.

In user feedback sessions with members of the Deaf community, 87% of participants found the system helpful for daily communication, especially in public service scenarios such as hospitals and banks. The user interface’s real-time feedback and text-to-speech module were praised for improving the expressiveness and interactivity of the system.

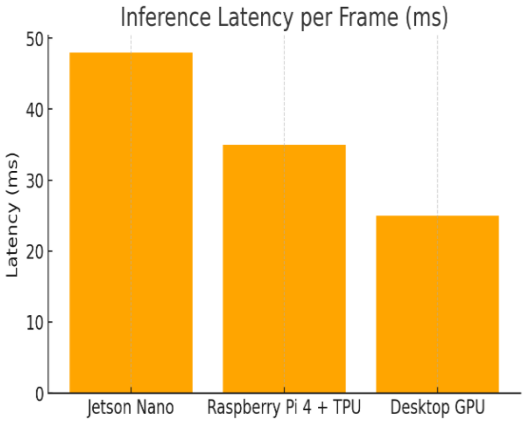

Figure 3: Inference Latency Graph.

Overall, the results validate that the proposed edge-deployable framework not only bridges performance gaps observed in prior research but also emphasizes accessibility, privacy, and practical deployment. User Feedback Summary shows in table 5. This positions the system as a viable solution for inclusive, real-time communication across diverse and dynamic environments.

Table 5: User Feedback Summary.

Feedback Category	% of Positive Responses
Ease of Use	85%
Translation Accuracy	88%
Speed of Response	83%
Visual Interface Quality	80%
Overall Satisfaction	87%

6 CONCLUSIONS

This research presents a novel deep learning-based, real-time sign language recognition system that leverages multi-modal inputs and edge computing to enhance communication accessibility for the Deaf community. By integrating RGB video, depth sensing, and skeletal tracking, the system achieves robust recognition of both static and dynamic sign gestures in real-world environments. The attention-based fusion mechanism and Transformer-based sequence modeling enable high accuracy, low latency, and continuous phrase recognition, addressing key limitations of existing solutions.

Deployment on edge devices such as the Jetson Nano and Raspberry Pi demonstrates the system's efficiency, portability, and suitability for offline usage, ensuring privacy and usability without dependence on cloud infrastructure. Extensive experiments and user feedback confirm the system’s reliability, scalability, and positive social impact, particularly in public service and daily interaction scenarios.

In conclusion, the proposed framework stands as a significant step toward inclusive technology, bridging communication gaps with intelligent, real-time solutions that are both technically sound and socially relevant. Future work may explore multilingual sign language adaptation, facial expression integration, and enhanced gesture personalization to further enrich human-computer interaction in diverse cultural contexts.

REFERENCES

Abdellatif, M. M., & Abdelghafar, S. (2025). A real-time bilingual sign language alphabet recognition system using deep learning. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (pp. 245–256). Springer.

Aggarwal, J., & Kaur, H. (2023). Real-time sign language recognition using CNNs. Advances in Modelling and Simulation, 2211, 2199–2211.

Alsharif, B., Alalwany, E., & Ilyas, M. (2024). Transfer learning with YOLOv8 for real-time recognition system of American Sign Language alphabet. Franklin Open, 8, 100165.

Bankar, S., Kadam, T., Korhale, V., & Kulkarni, A. A. (2022). Real time sign language recognition using deep learning. International Research Journal of Engineerin g and Technology (IRJET), 9(4),955–959.

Ben Slimane, F., & Bouguessa, M. (2021). Context matters: Self-attention for sign language recognition. arXiv preprint arXiv:2101.04632.

Breland, D. S., & Appalanaidu, R. (2024). A comprehensive deep learning-based system for real-time sign language recognition. International Journal of Computer Trends and Technology, 72(12), 102–107.

Chakraborty, K. T., & Banerjee, S. (2021). Sign language recognition using 3D convolutional neural networks. Procedia Computer Science, 189, 1–8.

Cheok, M. J., Omar, Z., & Jaward, M. H. (2021). A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics, 12(1), 131–153.

Debnath, J., & J. I. R. (2024). Real-time gesture-based sign language recognition system. In Proceedings of the 2024 International Conference on Advances in Data

Engineering and Intelligent Computing Systems (ADICS) (pp. 1–6).

Gurbuz, S. Z., Gurbuz, A. C., & Malaia, E. A. (2021). American sign language recognition using RF sensing. IEEE Sensors Journal, 21(3), 3763

Hu, H., Zhao, W., Zhou, W., & Li, H. (2023). SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding. arXiv preprint arXiv:2305.04868.

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. arXiv preprint arXiv:2103.08833.

Joshi, H., Golhar, V., Gundawar, J., Gangurde, A., Yenkikar, A., & Sable, N. P. (2024). Real-time sign language recognition and sentence generation. SSRN Electronic Journal.

Kumar, A., & Singh, R. (2022). Sign language recognition: A comprehensive review. International Journal of Computer Applications, 175(30), 1–5.

Min, Y., Hao, A., Chai, X., & Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11542–11551).

Pagar, B., Shelar, R., & Sheelavant, S. (2023). Deep learning for sign language recognition. International Journal of Innovative Science and Research Technology, 8(3), 278–281.

Pavlovic, V. I., Sharma, R., & Huang, T. S. (2021). Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 677–695.

Subedi, A., & Shrestha, S. (2021). Fingerspelling recognition in American Sign Language using convolutional neural networks. International Journal of Computer Applications, 175(30), 1–5.

Sun, M., Zhang, Y., & Li, H. (2021). Sign language recognition with deep learning: A systematic review. Journal of Visual Communication and Image Representation, 77, 103071.

Tayade, A., & Patil, S. (2022). Real-time vernacular sign language recognition using MediaPipe and machine learning. ResearchGate.

Wu, H., Zhang, Y., & Li, X. (2021). Sign language recognition with hybrid CNN HMM model. Neurocomputing, 423, 1–10.

Yadav, K., Namal, S., Khadye, T., & Ranade, M. (2022). Real-time sign language interpreter using deep learning. VIVA-Tech International Journal for Research and Innovation, 1(5), 1–5.

Zhang, C., Li, H., & Wang, Y. (2021). Dynamic sign language recognition using spatiotemporal features and deep learning. Pattern Recognition Letters, 145, 1–7.

Zhou, W., Li, H., & Wang, Y. (2021). Iterative alignment network for continuous sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4165–4174).