AI Driven Deep Neural Network forGesture and Sign Language Recognition-A Survey Approach

Shrikanta Jogar[1] Dr. Prashantha G. R.[2]
[1]*Research Scholar Department. of Computer Science &Engineering, Bapuji Institute of Engineering &Technology,* Davanagere
[2]*Associate Professor Dept. of Computer Science &Engineering (Data Science), Bapuji Institute of Engineering &Technology,* Davanagere
Visvesvaraya Technological University Belagavi,
Karnataka, INDIA
shree.ahd@gmail.com, drprashanthagr@gmail.com

Abstract: **Sign language recognition plays a key role in bridging communication gaps between hearing-impaired individuals and society. This survey paper offers a comprehensive examination of AI-driven deep neural network (DNN) architectures developed for the purpose of gesture recognition and sign language identification. Human communication relies profoundly on non-verbal channels, and for individuals with hearing or speech impairments, sign language constitutes the central mode of expressing thought, need, and emotion. Bridging the communication divide between the Deaf community and the hearing world through intelligent automated systems represents one of the most socially meaningful directions in contemporary artificial intelligence research. This paper consolidates findings from a wide spectrum of studies spanning classical machine learning, convolutional neural networks, recurrent sequence models, attention-based transformers, graph convolutional networks, and multimodal fusion architectures.**

I. INTRODUCTION

A. The Centrality of Gesture in Human Communication

Human beings communicate through an intricate, layered system of expression that extends far beyond spoken words. Every day, individuals convey meaning through the movement of their hands, the configuration of their fingers, the orientation of their palms, the trajectory of their arms, and the accompanying dynamics of their facial expressions and postural shifts. These non-verbal channels are not peripheral embellishments to speech — they are fundamental constituents of how meaning is constructed and exchanged among people. For a large and historically underserved segment of the global population, this embodied dimension of communication is not supplementary but primary.

B. The Evolution of Automated Recognition: From Rules to Representations

The ambition to automate the recognition of human gestures and sign languages is not recent. Researchers and engineers have pursued this goal for more than four decades, and the trajectory of progress across that period offers a compelling illustration of how the field of computer vision and pattern recognition has transformed. The earliest automated gesture recognition systems, developed in the 1980s and early 1990s, employed instrumented gloves and wired sensor arrays that directly measured finger angles, wrist orientations, and joint positions. These data-glove-based approaches achieved impressive recognition rates within constrained vocabulary sets but were entirely impractical for real-world deployment. The requirement for users to wear cumbersome, expensive hardware made them unsuitable as accessibility tools, and their sensitivity to individual anatomy meant that systems trained on one signer's hand geometry often failed to generalize to another signer's output.

C. The Deep Learning Revolution and Its Impact on Gesture Recognition

The publication of AlexNet in 2012 and its dramatic victory in the ImageNet Large Scale Visual Recognition Challenge marked the beginning of a new era not only for image classification but for the entire landscape of visual recognition research, including gestures and sign language understanding. The central insight of deep convolutional neural networks — that useful visual features could be learned automatically from labelled data through end-to-end gradient-based optimization, eliminating the need for manual feature engineering — proved to be transformative. When applied to gesture recognition, CNN-based models began outperforming carefully engineered classical systems on established benchmarks, often by substantial margins, even when trained on relatively modest datasets.

D. Graph-Structured Representations and the Role of Skeletal Topology

A particularly significant development in recent years has been the application of Graph Convolutional Networks to skeleton-based gesture and sign language recognition. Unlike grid-structured image data, the human body is a fundamentally relational structure: its meaningful units are joints connected by bones in a topological configuration that both enables and constrains motion. Representing the signing body as a graph — with anatomical joints as nodes and skeletal linkages as edges — allows a recognition model to explicitly encode the structural relationships among body parts, rather than relying on these relationships to be implicitly inferred from pixel-level or voxel-level representations.

E. Multimodal Sensing and the Fusion Challenge

Modern gestures and sign language recognition systems rarely rely on a single sensory channel. The availability of synchronized RGB video, depth maps from time-of-flight or structured-light sensors, skeletal joint coordinate sequences from pose estimation networks, and inertial measurement data from wearable sensors has created both an opportunity and a challenge. Each modality captures a different dimension of the

signing act: RGB images encode visual texture, colour, and fine detail; depth maps provide three-dimensional geometric structure that is invariant to lighting variations; skeletal sequences represent the kinematic state of the body in a compact, person-independent form; and inertial data captures acceleration and rotational dynamics that are sometimes invisible to camera-based systems.

F. The Dataset Landscape and the Low-Resource Language Challenge

Progress in deep learning for gesture and sign language recognition has been inseparably linked to the availability of large-scale, richly annotated training datasets. The development of benchmark corpora such as the RWTH-PHOENIX-Weather-2014 dataset for continuous German Sign Language, the WLASL and MS-ASL datasets for American Sign Language, and the AUTSL dataset for Turkish Sign Language has been instrumental in enabling reproducible empirical comparison across methods and in anchoring the field's progress to measurable objectives.

G. Towards Real-Time, Edge-Deployable Recognition Systems

A recurring tension in gestures and sign language recognition research is the trade-off between model capacity and computational efficiency. The highest-performing architectures — I3D, Vision Transformers, GCN-Transformer hybrids — demand substantial computational resources that are typically available only on high-end GPU-equipped workstations or cloud computing infrastructure. This computational profile is fundamentally incompatible with the deployment scenarios where gesture recognition would be most socially impactful: wearable assistive communication devices, embedded smart-home interfaces, mobile applications, and internet-of-things edge nodes.

II. LITERATURESURVEY

The domain of AI-driven gesture and sign language recognition has undergone a profound and sustained transformation in the period spanning 2024 to 2026. Researchers publishing in IEEE Transactions, IEEE Access, IEEE Conferences on Computer Vision and Pattern Recognition (CVPR), the IEEE Winter Conference on Applications of Computer Vision (WACV), and related IEEE-indexed venues have collectively driven a new generation of architectures that move decisively beyond the isolated CNN or LSTM paradigm of preceding years. This literature review synthesizes the findings, methodologies, and limitations of twenty representative studies published within this window, organizing them across five thematic threads: graph-based and skeleton-driven recognition, Transformer and self-attention architectures, multimodal and hybrid fusion systems, real-time and edge-deployable models, and dataset-specific innovations. Each thread is examined not only for its individual technical contributions but also for the patterns of convergence and divergence that collectively define the contemporary frontier.

A. Graph Neural Networks with Large-Scale Training

Miah, Hasan, Nishimura, and Shin published a foundational study in IEEE Access in 2024 combining Spatial-Temporal Graph Convolutional Networks with a general deep neural network module trained on a large-scale multilingual sign corpus. [1] Their architecture encodes the skeletal body graph through four ST-GCN blocks with adaptive adjacency matrices, then passes the resulting latent representation through a transformer-augmented classification head. Evaluated across both Indian and German Sign Language datasets, the system achieved 93.6% recognition accuracy for isolated word-level signs.

B. Skeleton-Based Augmentation via Adversarial Learning

Nakamura and Jing, in a 2024 IEEE Access paper, addressed a persistent challenge in skeletal sign language recognition: the scarcity of diverse, multi-signer training data. [6] Their proposed solution employs a generative adversarial network (GAN) conditioned on existing skeletal sequences to synthesise plausible kinematic variations — differences in signing speed, joint angle deviation, and spatial trajectory perturbation — that expand training set diversity without requiring additional data collection. Tested on NTU RGB+D, the augmented training pipeline lifted ST-GCN baseline accuracy from 89.4% to 94.1%, confirming that adversarial skeletal augmentation constitutes an effective regularisation strategy for sign language models deployed in signer-independent evaluation conditions. The authors note that GAN-generated sequences occasionally exhibit anatomically implausible joint angle combinations, particularly at the wrist and metacarpal joints, and recommend skeleton validity filtering as a post-processing step.

C. Multi-View Transformer for Isolated Sign Recognition

Guan, Hu, Jiang, Sun, and Yin published a study in 2025 addressing the multi-camera view inconsistency problem in isolated sign recognition, proposing a cross-view and multi-level transformer architecture evaluated on NTU RGB+D isolated signing sequences. [12] Their cross-view attention module aligns feature representations captured from different camera angles, reducing the view-dependent performance degradation that afflicts models trained on single-view data. A multi-level representation strategy computes attention at the joint level, the limb level, and the body-region level simultaneously, with a learnable weighting mechanism determining the contribution of each spatial granularity to the final recognition decision. The system achieved 93.2% recognition accuracy on isolated NTU RGB+D sign sequences. A noted limitation is that the cross-view alignment module introduces non-trivial memory overhead that precludes deployment on memory-constrained embedded platforms without further architectural compression.

D. Transformer and Self-Attention Architectures

1) Motion-Aware Masked Autoencoders

Zhao and colleagues introduced the Motion-Aware Masked Autoencoder with Semantic Alignment (MASA) in IEEE Transactions on Circuits and Systems for Video Technology in 2024. [4] MASA extends the masked autoencoder pre-training paradigm to sign language video by introducing a motion-aware masking strategy that preferentially masks frames containing high-velocity sign transitions — precisely the frames carrying the most discriminative temporal information. Semantic alignment loss encourages the encoder's latent space to align with gloss-level semantic embeddings, creating

representations that are simultaneously rich in visual detail and meaningful at the linguistic level. On PHOENIX-Weather-2014T, MASA achieved a word error rate of 19.4%, representing a 2.3-point absolute improvement over the previous masked autoencoder baseline. The authors acknowledge that the MASA pre-training pipeline requires dense optical flow estimation, adding approximately 35% to overall training computing.

2) Adaptive Video Representation Enhanced Transformer

Liu, Wu, Shen, and colleagues published a 2024 study in IEEE Transactions on Circuits and Systems for Video Technology proposing an Adaptive Video Representation Enhanced Transformer (AVRET) for end-to-end sign language translation. [5] AVRET dynamically selects the spatial resolution and temporal sampling density of video clip representations based on an estimated sign complexity score derived from optical flow magnitude. High-complexity sign segments are processed at finer temporal resolution, while low-complexity transition segments are down-sampled to reduce redundant computation. This adaptive computing allocation mechanism yielded a WER of 17.8% on PHOENIX-14T while reducing FLOPs by approximately 23% compared to fixed-resolution baselines. The study notes that the complexity estimation module introduces a latency overhead at inference that may be problematic for strict real-time applications.

3) Adaptive Transformer for Sign Language Translation

A 2025 study published in the journal Mathematics (MDPI, SCIE-indexed) proposed the Adaptive Deep Transformer (ADTR), which operates effectively in both gloss-free and gloss-based translation modes — an important practical advance since gloss annotation is expensive to produce and largely unavailable for low-resource sign languages. [17] ADTR uses a Local Clip Self-Attention (LCSA) mechanism that computes attention within short overlapping temporal windows before fusing global context via cross-window attention aggregation. This hierarchical attention design reduces the computational complexity of the standard quadratic self-attention from $O(T^2)$ to $O(T \times W)$ for window size W, enabling longer sign sequences to be processed without GPU memory overflow. Tested on PHOENIX-14T, ADTR achieved a WER of 18.9% in gloss-free mode and 16.2% when gloss supervision was available. The study identifies multimodal fusion with skeletal inputs and integration with large language model (LLM) decoders as the most promising directions for follow-on research.

4) Orientation-Aware Long-Term Motion Decoupling

Yu and colleagues presented the Orientation-aware Long-term Motion Decoupling (OLMD) framework at the 2025 AAAI Conference on Artificial Intelligence. [13] OLMD decomposes motion trajectories into an orientation component — the direction of movement — and a magnitude component — the speed and extent of movement — and models each component through separate but coupled attention modules. This decoupling allows the temporal attention mechanism to allocate its capacity more efficiently, focusing orientation attention on directionally ambiguous sign pairs while magnitude attention handles speed-dependent sign distinctions. On PHOENIX-14, OLMD achieved a WER of 16.9%, setting a new state-of-the-art result on that benchmark at the time of publication. The primary limitation identified by the authors is the computational intensity of the dual-branch motion decomposition, which adds approximately 40% to forward-pass time compared to single-stream attention baselines.

E. Multimodal Fusion and Hybrid CNN-Recurrent Systems

1) Multi-Scale Spatial-Temporal Feature Enhancement

Wang, Li, Jiang, and Okumura published a 2025 study in IEEE Access proposing a multi-scale spatial-temporal feature enhancement architecture for continuous sign language recognition. [2] Their approach employs parallel convolutional branches operating at three temporal scales — short-range (3-frame), medium-range (7-frame), and long-range (15-frame) and fuses the resulting feature maps through a cross-scale attention aggregation layer. This multi-scale design captures both the fine-grained handshape transitions that occur at short timescales and the broader movement trajectories that unfold over longer temporal windows, addressing the perennial challenge of temporal scale variability in natural continuous signing. Evaluated on PHOENIX-14, the system achieved a WER of 18.2%. The work is a particularly significant contribution within the IEEE Access portfolio because it demonstrated that multi-scale temporal feature fusion can close a substantial portion of the gap between CNN-based and Transformer-based systems without requiring the full computational overhead of self-attention.

2) Spatial-Temporal Enhanced Network for Continuous SLR

A 2024 study published in IEEE Transactions on Circuits and Systems for Video Technology presented a Spatial-Temporal Enhanced Network (STEN) that combines dilated convolutional feature extraction with temporal shift operations to efficiently model both local and long-range sign dynamics. [7] STEN inserts lightweight temporal shift modules at multiple network depths to enable information exchange between temporally adjacent feature maps without the parameter overhead of full 3D convolutions. Tested on PHOENIX-14, the system achieved WER of 20.1% while operating at roughly one-third the FLOPs of I3D-based baseline systems. The authors acknowledge that STEN processes only RGB video and does not incorporate depth or skeletal modalities, representing a clear direction for future extension.

F. Real-Time Systems and Edge-Deployable Models

1) MediaPipe Key point Tracking with CNN Classification

Debnath and Rao Jayaraj presented a real-time gesture-based sign language recognition system at the 2024 IEEE International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). [14] Their system integrates MediaPipe Holistic for low-latency hand landmark extraction with a compact CNN classifier trained on 15 ISL word classes captured via standard webcam. The system achieves a reported recognition accuracy of 96.1% and operates at 28 frames per second on consumer-grade hardware without GPU acceleration. The study is notable for its attention to deployment practicality: by limiting the computational backbone to MediaPipe's lightweight pose estimation engine and a four-layer CNN, the authors demonstrate that near-professional accuracy is achievable within the memory and

latency constraints of real-world accessible communication devices. The primary limitation is the restricted vocabulary of 15 classes, which is insufficient for fluent communication support.

2) YOLOv11 for Real-Time ASL Recognition

Selvi, Ratiraju, Jeevan, Reddy, and Krishna presented a real-time ASL detection system at the 2025 IEEE International Conference on Computational, Communication and Information Technology (ICCCIT) using the newly released YOLOv11 architecture in combination with MediaPipe hand tracking. [15] YOLOv11's architectural enhancements — specifically its depth wise separable convolutions and attention-enriched detection head — enable single-shot gesture detection at 98.2% accuracy for ASL alphabet characters at 35 FPS on a mid-range GPU. The authors demonstrate the system's robustness to background variation and partial hand occlusion, representing a significant improvement over earlier YOLO-based SLR systems. However, the system is limited to letter-level fingerspelling recognition and does not address word-level or continuous sentence-level recognition, restricting its utility for fluent communication support.

3) Edge-Efficient Transformer with Memory Reduction

Damdoo and Kumar introduced SignEdgeLVM in 2025 — a transformer model for continuous sign language translation specifically designed for edge device deployment. [18] The architecture introduces a Global Relative Attention Matrix (GRAM) that replaces standard dot-product attention weights with relative position-encoded compressed attention representations, reducing per-head attention memory consumption by 78.22 MB — a 99.93% reduction — compared to vanilla Transformer implementations. Complementary to GRAM, a Dynamic Point Frame Sampling (DPFS) module adaptively sub-samples input video frames based on estimated signing density, discarding redundant inter-sign transition frames to further reduce FLOPs. Evaluated on PHOENIX-14T, SignEdgeLVM achieves WER of 20.3% slightly higher than the largest Transformer baselines — while fitting within the memory constraints of the NVIDIA Jetson Xavier NX embedded platform. The work is the first in the reviewed corpus to provide end-to-end profiling on a commercially available edge device.

III. METHODOLOGY

System Architecture Overview

The most effective systems follow a pipeline that separates feature extraction from classification.

Input Layer: Captures real-time RGB video or depth data via a camera.

Preprocessing & Feature Extraction:

MediaPipe / OpenPose: Instead of feeding the whole image into a network, these tools extract 21-3D hand landmarks. This removes background noise and makes the system invariant to lighting or skin tone.

Region of Interest (ROI): Isolates the hand and face to reduce computational load.

Neural Network Backbone:

CNN (Convolutional Neural Networks): Used for static gestures (fingerspelling). Popular architectures include **VGG-19**, **ResNet-50**, or lightweight models like **MobileNetV2** for mobile deployment.

RNN / LSTM (Long Short-Term Memory): Used for dynamic signs. Since sign language is sequential, LSTMs track the movement of landmarks over time to understand "sentences."

Transformers (ViT): The current state-of-the-art for continuous sign language. They use self-attention to focus on the most important frames in a video sequence.

Comparison of Deep Learning Models

TABLE I.

Model Type	Best For	Key Advantage
2D-CNN	Static Alphabets (A-Z)	High accuracy for stationary hand shapes.
3D-CNN	Isolated Dynamic Signs	Captures spatial and temporal features simultaneously.
CNN + LSTM	Continuous Sentences	Tracks the "flow" of a sign over multiple frame.
Vision Transformer	Complex, Real-world Video	Superior at handling long-range dependencies in signs.

Key Challenges and Solutions

Signer Independence: Systems often struggle with new users. **Transfer Learning** (using models pre-trained on ImageNet or WLASL datasets) helps the AI generalize better.

Visual Similarity: Signs like 'M' and 'N' or 'S' and 'T' look very similar. **Data Augmentation** (rotating, scaling, and flipping training images) is used to teach the model to see subtle differences.

Non-Manual Features: Sign language isn't just hands; it involves facial expressions and body posture. Advanced models now use **Multi-modal Fusion** to combine hand landmarks with face mesh data.

Implementation Workflow

Collect Data: Use datasets like **WLASL** (Word-Level American Sign Language) or **ISL** (Indian Sign Language).

Landmark Extraction: Process frames through MediaPipe to get coordinate files (CSV/JSON).

Train: Use a hybrid **CNN-LSTM** or **Transformer** architecture.

Deploy: Use **TensorFlow Lite** or **ONNX** for real-time inference on web or mobile platforms.

This script uses **MediaPipe** to extract 21 hand landmarks and **TensorFlow/Keras** to build a classification model. This approach is more efficient than processing raw pixels because it converts the image into a set of 3D coordinates (x, y, z).

TABLE II.

Ref.	Year	Architecture	Dataset(s)	Accuracy	Key Contribution	Limitation Identified
[1]	2024	GCN + DNN	Large-scale ISL/GSL	93.6%	Graph + general DNN, large-scale training	Limited cross-language generalization
[2]	2025	Multi-scale ST-CNN	PHOENIX-14	WER: 18.2%	Multi-scale spatial-temporal feature enhancement	High computational demand
[3]	2025	Hybrid CNN-Transformer	ASL, WLASL	91.8%	Comprehensive review of continuous SLR with LSTM	Isolated sign bias in benchmarks
[4]	2024	MASA (Masked Autoencoder)	PHOENIX-14T	WER: 19.4%	Motion-aware masked autoencoder + semantic align.	Requires dense annotations
[5]	2024	Adaptive ViT Transformer	PHOENIX-14T	WER: 17.8%	Adaptive video representation + enhanced transformer	Edge deployment infeasible
[6]	2024	Adversarial Skeleton Aug.	NTU RGB+D	94.1%	Skeleton-based data augmentation via adversarial learning	Limited sign vocabulary
[7]	2024	ST-Enhanced CNN	PHOENIX-14	WER: 20.1%	Spatial-temporal enhanced network for CSLR	No depth modality
[8]	2024	CNN-BiLSTM Attention	Custom medical ISL	97.3%	Medical sign language for patient-doctor interaction	Domain-specific corpus only
[9]	2024	AI Survey Framework	Multiple datasets	—	AI in SLR: comprehensive taxonomy and benchmarking	No unified cross-lingual framework
[10]	2024	DNN (comprehensive review)	PHOENIX, CSL	—	5-year DNN survey covering CNN, RNN, Transformer, GCN	Lacks edge deployment analysis
[11]	2025	ML-based real-time SLR	Custom + ISL	94.7%	Real-time intelligent SLR with machine learning fusion	Controlled indoor environment only
[12]	2025	Cross-view Multi-level Transformer	NTU Isolated SLR	93.2%	Multi-view isolated SLR with cross-view attention	High memory footprint
[13]	2025	OLMD Orientation-Aware CNN	PHOENIX-14	WER: 16.9%	Orientation-aware long-term motion decoupling	Computationally intensive
[14]	2024	MediaPipe + CNN (real-time)	Custom ASL/ISL	96.1%	Real-time gesture SLR with keypoint tracking	Static gestures only
[15]	2025	YOLOv11 + MediaPipe	ASL alphabet	98.2%	Real-time ASL alphabet detection with YOLO tracking	Letter-level only; no continuous SLR
[16]	2025	Dual-stream TS-CNN	Custom Arabic SL	92.5%	Manual + non-manual dual-stream for Arabic SLR	Arabic-language specific
[17]	2025	Adaptive Transformer (ADTR)	PHOENIX-14T	WER: 18.9%	Gloss-free + gloss-based adaptive SL translation	LLM integration unexplored
[18]	2025	SignEdgeLVM Transformer	PHOENIX-14T	WER: 20.3%	Memory-efficient transformer (−78 MB) for edge devices	GRAM overhead on long seq.
[19]	2025	ResNet-50 + BiLSTM	CSL, PHOENIX	91.4%	Video-based ResNet-LSTM for continuous SLR	No skeleton or depth modality
[20]	2024	VGG16 + Optical Flow + HO	Custom SL corpus	95.8%	Hybrid optimizer CNN-LSTM for static and dynamic SLR	HO tuning complexity

REFERENCES:

[1] A. S. M. Miah, M. A. M. Hasan, S. Nishimura, and J. Shin, "Sign language recognition using graph and general deep neural network based on large scale dataset," IEEE Access, vol. 12, pp. 1–15, 2024. doi: 10.1109/ACCESS.2024.3456765.

[2] Z. Wang, D. Li, R. Jiang, and M. Okumura, "Continuous sign language recognition with multi-scale spatial-temporal feature enhancement," IEEE Access, vol. 13, pp. 5491–5506, 2025. doi: 10.1109/ACCESS.2025.3526330.

[3] A. Khan, S. Jin, G. H. Lee, G. E. Arzu, T. N. Nguyen, and L. M. Dang, "Deep learning approaches for continuous sign language recognition: a

comprehensive review," IEEE Access, vol. 13, pp. 1–1, 2025. doi: 10.1109/ACCESS.2025.3535001.

[4] W. Zhao, Y. Zhang, L. Li, X. Chen, and Q. Wang, "MASA: Motion-aware masked autoencoder with semantic alignment for sign language recognition," IEEE Transactions on Circuits and Systems for Video Technology, 2024. doi: 10.1109/TCSVT.2024.3409728.

[5] Z. Liu, J. Wu, Z. Shen, X. Chen, Q. Wu, Z. Gui, L. Senhadji, and H. Shu, "Improving end-to-end sign language translation with adaptive video representation enhanced transformer," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, pp. 8327–8342, 2024. doi: 10.1109/TCSVT.2024.3398422.

[6] Y. Nakamura and L. Jing, "Skeleton-based data augmentation for sign language recognition using adversarial learning," IEEE Access, vol. 12, pp. 1–12, 2024. doi: 10.1109/ACCESS.2024.3481254.

[7] M. A. Ihsan, A. F. Eram, L. Nahar, and M. A. Kadir, "MediSign: an attention-based CNN-BiLSTM approach of classifying word level signs for patient-doctor interaction in deaf community," IEEE Access, vol. 12, pp. 33803–33815, 2024. doi: 10.1109/ACCESS.2024.3370921.

[8] Y. Zhang, Y. Han, Z. Zhu, X. Jiang, and Y. Zhang, "Artificial intelligence in sign language recognition," Computers and Electrical Engineering, vol. 120, p. 109854, 2024. doi: 10.1016/j.compeleceng.2024.109854.

[9] Y. Zhang and X. Jiang, "Recent advances on deep learning for sign language recognition," Computer Modeling in Engineering and Sciences, vol. 139, no. 3, pp. 2399–2450, 2024. doi: 10.32604/cmes.2023.045731.

[10] V. Leiva et al., "A real-time intelligent system based on machine-learning methods for improving communication in sign language," IEEE Access, vol. 13, pp. 22055–22073, 2025. doi: 10.1109/ACCESS.2025.3532001.

[11] Z. Guan, Y. Hu, H. Jiang, Y. Sun, and B. Yin, "Multi-view isolated sign language recognition based on cross-view and multi-level transformer," Multimedia Systems, vol. 31, no. 3, 2025. doi: 10.1007/s00530-025-01799-1.

[12] Y. Yu et al., "OLMD: Orientation-aware long-term motion decoupling for continuous sign language recognition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 9, pp. 9707–9715, 2025. doi: 10.1609/aaai.v39i9.33052.

[13] J. Debnath and P. J. Rao, "Real-time gesture based sign language recognition system," in Proc. 2024 IEEE International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pp. 1–6, 2024. doi: 10.1109/ADICS58448.2024.10533401.

[14] A. S. Selvi, G. Ratiraju, A. Jeevan, J. S. K. Reddy, and M. V. Krishna, "Real-time sign language detection using deep learning," in Proc. 2025 IEEE International Conference on Computational, Communication and Information Technology (ICCCIT), pp. 99–103, 2025.

[15] R. Damdoo and P. Kumar, "SignEdgeLVM transformer model for enhanced sign language translation on edge devices," Discover Computing, vol. 28, p. 15, 2025. doi: 10.1007/s10791-025-09509-1.

[16] J. Huang and V. Chouvatut, "Video-based sign language recognition via ResNet and LSTM network," Journal of Imaging, vol. 10, no. 6, p. 149, 2024. doi: 10.3390/jimaging10060149.