IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(RESEARCH ARTICLE)

Sign language recognition in the deep learning era: A comprehensive study of model performance, robustness and deployment considerations

Ashish Kumar Walter [1], and Garima Srivastava [1, *] and Lalita Kumari [2]

[1] *Department of Engineering and Technology Amity University Lucknow, India.*
[2] *Department of Engineering and Technology Amity University Patna, India.*

International Journal of Science and Research Archive, 2025, 15(03), 398-407

Publication history: Received on 25 April 2025; revised on 04 June 2025; accepted on 06 June 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.3.1699

Abstract

This paper presents a dual-domain evaluation of classical and modern architectures for Sign Language Recognition (SLR) and Traffic Sign Classification (TSC), addressing critical challenges in accessibility and autonomous systems. We conduct a comprehensive assessment of 20 SLR models, spanning CNNs, hybrid CNN-LSTM pipelines, and transformer-based frameworks, evaluated on the Sign Language MNIST and ASL Fingerspelling datasets. Performance is measured across accuracy, computational efficiency, and robustness metrics.

For TSC, we benchmark 10 models—including lightweight CNNs, vision transformers, and object detectors—on the GTSRB, BelgiumTS, and TT100K datasets. The study examines classification and detection performance under varying noise conditions to assess real-world applicability. We analyze trade-offs between model complexity, inference speed, and deployment feasibility, providing guidelines for edge-optimized implementations.

Keywords: Sign Language Recognition (SLR); Traffic Sign Classification (TSC); Deep Learning Architectures; Computational Efficiency; Edge Deployment

1. Introduction

In recent years, computer vision has rapidly matured from a laboratory curiosity to a core enabler of intelligent systems across various domains. Among the most socially impactful applications of this progress are Sign Language Recognition (SLR) and Traffic Sign Classification (TSC)—two fields that, while distinct in their end goals, share critical technical challenges such as pattern recognition under visual variability, real-time inference demands, and robustness to real-world noise. SLR plays a vital role in improving accessibility for the deaf and hard-of-hearing communities by enabling automatic translation of sign language into text or speech [1]. TSC, meanwhile, is a cornerstone of Advanced Driver Assistance Systems (ADAS) and autonomous driving technologies, where accurate interpretation of road signs is essential for safety and navigation.

The practical success of SLR and TSC systems depends not only on high recognition accuracy but also on their ability to operate under variable lighting, backgrounds, occlusions, motion blur, and signer or sign diversity. Traditional rule-based approaches and shallow classifiers often fail in such unconstrained environments, leading researchers to adopt deep learning architectures—particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) models, and more recently, Transformers and Graph Neural Networks (GNNs).

Sign Language Recognition presents unique challenges: unlike spoken language, sign language is a multimodal, spatial-temporal visual language, involving hand shapes, movements, facial expressions, and body posture. Capturing this complexity requires models capable of extracting both static and dynamic features while maintaining real-time

[*] Corresponding author: Garima Srivastava

performance. Recent advancements have explored hybrid models (e.g., CNN-LSTM), depth-sensor fusion, pose estimation via OpenPose, and wearable motion capture devices. Meanwhile, the integration of Natural Language Processing (NLP) techniques has enabled some systems to go beyond gesture recognition to full sign-to-sentence translation, enhancing their utility in real-world communication.

On the other hand, Traffic Sign Classification requires models to operate with extremely low latency and high accuracy in dynamic and often unpredictable driving environments. The variability in sign appearance due to weather, lighting, occlusion, and geographical diversity poses significant challenges. Moreover, the demand for deployment on edge devices like car dashboards or mobile processors necessitates models with small footprints, minimal energy consumption, and fast inference speeds [2]. State-of-the-art TSC systems now leverage lightweight architectures such as MobileNet, EfficientNet, ShuffleNet, and transformer-based variants like E-MobileViT, in addition to classic object detectors like YOLOv5.

Despite significant progress, both SLR and TSC systems remain susceptible to real-world distortions. Models that perform well in controlled laboratory conditions often degrade significantly under environmental noise or cross-dataset domain shifts. This underscores the need for rigorous evaluations not only of accuracy but also of robustness, generalizability, and deployment readiness.

This study is motivated by the need for a comprehensive, cross-cutting analysis of the current state of SLR and TSC systems. Our contribution is fourfold:

- **Breadth of Evaluation:** We survey and analyze 30 cutting-edge deep learning architectures used in SLR and TSC, covering CNNs, RNNs, transformers, hybrid models, and GNNs.
- **Depth of Analysis:** We assess models not only by accuracy but also by their ability to generalize across datasets and maintain performance under controlled perturbations such as noise, occlusion, and motion blur.
- **Resource Profiling:** We provide a detailed resource utilization profile—including inference latency, model size, FLOPs, and memory footprint—across both high-capacity and edge-optimized models.
- **Deployment Considerations:** We discuss the feasibility of real-world implementation, including guidelines for selecting models for embedded or low-power environments, and highlight critical limitations and areas for future research.

By combining literature synthesis with empirical performance analysis on diverse datasets—including Sign Language MNIST, ASL Fingerspelling Video, GTSRB, BelgiumTS, and TT100K—this work presents a unified perspective on the dual challenges and innovations within SLR and TSC. Our findings not only showcase the promise of emerging architectures like transformers and edge-optimized CNNs, but also stress the importance of data diversity, augmentation strategies, multimodal input, and model robustness for translating laboratory success into practical deployments.

Ultimately, the goal of this research is to inform and guide the development of next-generation intelligent systems that are not only accurate, but also robust, efficient, and accessible—whether they are deployed in a car navigating busy streets or in a mobile app enabling inclusive communication for the hearing-impaired.

2. Literature Survey

This literature survey highlights the advancements and challenges in Sign Language Recognition. While deep learning and multimodal techniques significantly improve accuracy [2], real-time deployment, continuous sign recognition, and dataset quality remain challenges. Future work should focus on optimizing models for mobile use, enhancing sign transition handling, and leveraging AI for real-time translation.

Table 1 Literature Survey

Citation No.	Methods	Finding
3	CNN and RNN models with transfer learning on large datasets	Achieved 90% accuracy in real-time settings with effective deep learning feature extraction.
4	LSTM architecture with optical flow tracking and sequence-based modelling	Achieved 85% accuracy for continuous recognition but overlapping gestures remained challenging.
5	CNN + LSTM hybrid model with depth sensors and multi-feature extraction	Achieved 92% accuracy with multimodal fusion, demonstrating enhanced robustness for dynamic gestures.
6	MobileNet-based deep learning with hand tracking algorithms	Achieved 88% accuracy on mobile devices with low computational overhead suitable for real-time applications.
7	Real-time sign translation system with NLP techniques for gesture mapping	Achieved 87% accuracy for phrase-level recognition with improved usability through NLP integration.
8	Wearable motion sensors with deep learning classifiers and edge AI processing	Achieved 91% accuracy with edge computing optimization but faced challenges with sensor drift and calibration.
9	OpenPose for key point detection with RNN for sequence classification	Achieved 89% accuracy effective for dynamic gestures but occlusion remained a challenge.
10	CNN + SVM hybrid approach with feature extraction from images and motion	Achieved 86% accuracy with improved robustness through hybrid methods but dependent on high-quality input.
11	Performance benchmarking using CNN and RNN with large dataset training	CNN outperformed RNN in static gestures while RNNs were more efficient for sequential gestures.
12	Transformer + TTS model with context-aware NLP integration	Achieved 88% accuracy with feasible real-time conversion but struggled with complex grammar.
13	Pre-trained CNN models fine-tuned for SLR with data augmentation	Achieved 90% accuracy on limited data with effective training time reduction through transfer learning.
14	3D CNN with depth estimation and multi-frame sequence analysis	Achieved 91% accuracy effective for dynamic gestures but struggled with fast motion recognition.
15	3D CNNs for depth-based recognition with motion tracking algorithms	Achieved 90% accuracy in low-light scenarios with enhanced tracking in occluded environments.
16	Transformer networks with attention mechanisms and data augmentation	Achieved 93% accuracy with enhanced sequential gesture recognition outperforming CNN-RNN approaches.
17	CNN for feature extraction with LSTM for temporal sequence analysis	Achieved 89% accuracy with improved sequential recognition and reduced computational cost.
18	YOLO-based object detection with hand tracking algorithms	Achieved 87% accuracy with low-latency recognition but struggled with overlapping gestures.
19	Graph neural networks for relational modelling with pose estimation	Achieved 89% accuracy effective for large datasets with improved contextual understanding.
20	EfficientNet-based model compression with quantization for edge AI	Achieved 85% accuracy suitable for embedded systems with trade-off between speed and accuracy.
21	Optical flow tracking with motion-based gesture segmentation	Achieved 88% accuracy with better movement detection but struggled with fast hand motions.
22	Fine-tuned CNN models with data augmentation using synthetic gestures	Achieved 92% accuracy for ASL with effective regional variation handling through transfer learning.

This survey encompasses a diverse range of methodologies applied to sign language recognition (SLR), highlighting the progression from traditional CNN and RNN models to more advanced transformer architectures and edge AI solutions.

Across the studies, deep learning techniques consistently achieved high recognition accuracies—ranging from 85% to 93%—demonstrating their effectiveness in both static and dynamic gesture scenarios.

Hybrid approaches such as CNN-LSTM combinations [5,7] and the integration of multimodal inputs (e.g., depth sensors, motion tracking) notably improved recognition robustness and sequential modeling capabilities. Lightweight models like MobileNet [6] and EfficientNet [20] enabled real-time recognition on mobile and embedded platforms, emphasizing low latency and computational efficiency.

Several studies [7,12,16] integrated NLP and attention mechanisms, pushing SLR beyond gesture recognition into meaningful language translation with context awareness. Edge computing and wearable sensor-based approaches [8] also demonstrated promise, particularly for continuous, on-device recognition tasks.

Despite these advancements, challenges remain with overlapping gestures, occlusion, fast motion, and complex grammar. Innovations such as transformer networks [16] and graph neural networks [19] showed superior performance by enhancing contextual understanding and relational modeling.

In conclusion, the field is rapidly evolving toward more accurate, efficient, and context-aware SLR systems, with transformer-based and hybrid deep learning models leading current performance benchmarks.

3. Methodology

3.1. Data and Preprocessing

3.1.1. Sign Language Recognition (SLR) Datasets

Two major datasets were used for training and evaluation in the SLR domain:

- **Sign Language MNIST**: A static gesture dataset consisting of 27,455 grayscale images representing 24 alphabetic signs (excluding J and Z, which require motion).
- **ASL Fingerspelling Video Dataset**: A dynamic dataset comprising approximately 200,000 annotated video frames, each depicting fingerspelled letters and continuous sign sequences from multiple signers in varying lighting and backgrounds [25].

To standardize and optimize data quality for model training, the following preprocessing pipeline was applied:

- **Resizing**: All images and frames were resized to 28×28 pixels (SL MNIST) and 64×64 pixels (ASL Video) to reduce computational complexity while preserving essential spatial features.
- **Normalization**: Pixel intensity values were scaled to a (0, 1 range to stabilize training and accelerate convergence by preventing saturation of activation functions.

Augmentation Techniques

- **Random cropping** (±10%) and **rotation** (±15°) were used to introduce spatial variability and simulate viewpoint changes.
- **Brightness jitter** (±20%) allowed the model to generalize better under varied lighting conditions.
- **Horizontal flipping** was selectively applied, particularly useful for symmetric gestures, to expand dataset diversity.
- **Elastic distortions** were also tested to mimic subtle finger articulation differences between users.
- **Frame Sampling (for dynamic sequences)**: For video-based input, frames were uniformly sampled to capture gesture evolution across time while avoiding redundancy. Temporal smoothing was applied to reduce frame-level noise.

These steps collectively ensured that the models could learn from both canonical and real-world variants of each sign, enhancing generalization to different signers and contexts.

3.1.2. Traffic Sign Classification (TSC) Datasets

Three large-scale datasets were utilized for traffic sign classification and detection:

- **GTSRB (German Traffic Sign Recognition Benchmark)**: Comprises 39,209 training and 12,630 test images, covering 43 traffic sign classes.
- **BelgiumTS**: Contains 7,095 images from 62 traffic sign classes, used primarily for cross-dataset generalization testing.
- **TT100K**: A diverse dataset with 100,000 annotated images, used for object detection evaluation across complex road scenes.

To prepare these images for classification and detection models, the following preprocessing strategies were adopted:

Resizing

- **Classification models** received 64×64 resized inputs.
- **Detection models** used 640×640 resolution to maintain sufficient detail for bounding box regression.
- **Normalization**: RGB channels were standardized to zero mean and unit variance, which improves convergence behavior and is particularly beneficial for deeper CNNs.

Augmentation Techniques

- **CutMix** and **MixUp**: These regularization strategies combined multiple images and labels, helping the model learn more generalizable features.
- **Gaussian blur ($\sigma \leq 1.5$)** and **motion blur filters** were applied to simulate poor-quality road camera feeds or rapid vehicle movement.
- **Weather simulation filters** (e.g., fog, rain) were used to artificially inject noise reflective of adverse driving conditions.
- **Cutout** augmentation (random masks covering 15% of the image) mimicked partial occlusions from obstacles like tree branches or pedestrians.
- **Color Jitter and Contrast Stretching**: These techniques simulated varying lighting conditions such as glare, dusk, and shadows, improving model robustness to environmental shifts.

By carefully tailoring the preprocessing techniques to the specific demands of each domain, the training pipeline was designed not only to enhance model performance but also to bridge the domain gap between curated datasets and unpredictable real-world scenarios. The integration of domain-aware augmentations ensured that both SLR and TSC models could generalize effectively across unseen conditions, paving the way for reliable deployment in practical applications.

3.2. Training Protocol

- **Frameworks:** PyTorch 2.0 with CUDA acceleration; ONNX for cross-platform deployment.
- **Optimizers and Schedulers:** AdamW with 5-epoch linear warmup, cosine decay over 100 epochs for classification; SGD with momentum=0.9 and step decay for detection.
- **Hyperparameter Search:** Bayesian optimization (Optuna) over learning rates ((1e-5, 1e-3)), weight decays ((1e-6, 1e-3)), dropout ((0.1, 0.5)).
- **Early Stopping:** Monitored validation loss (patience=10, min Δ=0.001) and checkpointing at peak ROC-AUC.

Table 2 Evaluation Metrics

Metric	Definition
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-score	2 × Precision × Recall / (Precision + Recall)
ROC-AUC	Area under ROC curve (threshold-invariant)
mAP@50	Mean average precision at IoU ≥ 0.5 (detection, TT100K)
Latency (ms)	Inference time on ARM Cortex-A72 (edge) and NVIDIA V100 (server)
Model Size (MB)	Serialized model checkpoint size
FLOPs (G)	Multiply–accumulate operations per inference

4. Results and Discussion

This section presents the performance results of the 20 algorithms on the Sign Language MNIST dataset. The results are presented in tabular and graphical formats, showing the values of the evaluation metrics for each algorithm. A detailed analysis of the results is provided, comparing the performance of different algorithms and highlighting any significant findings. The performance of our YOLOv5 model [26] is analyzed within the context of these results.

Table 3 Performance Comparison

Algorithm	Accuracy	Precision	Recall	F1-score
LeNet-5	0.85	0.86	0.84	0.85
AlexNet	0.92	0.93	0.91	0.92
VGGNet	0.94	0.95	0.93	0.94
ResNet	0.96	0.97	0.95	0.96
Inception	0.95	0.96	0.94	0.95
MobileNet	0.91	0.92	0.90	0.91
RNN	0.88	0.89	0.87	0.88
LSTM	0.90	0.91	0.89	0.90
HOG + SVM	0.78	0.79	0.77	0.78
SIFT + SVM	0.82	0.83	0.81	0.82
k-NN	0.75	0.76	0.74	0.75
YOLOv5	0.98	0.983	0.986	0.984

The results show that deep learning approaches, particularly ResNet [27] and the YOLOv5 model, achieve the highest accuracy. Traditional methods generally perform less well, highlighting deep learning's advantages for complex pattern recognition. The YOLOv5 model's high precision and recall indicate a robust ability to classify sign language gestures. A detailed analysis of the confusion matrix for each algorithm would further reveal strengths and weaknesses and identify challenging gestures to classify. Statistical significance testing would be needed to confirm the significance of performance differences.

Table 4 SLR Performance Summary

Model	Acc (%)	F1 (%)	ROC-AUC	Latency (ms)	Size (MB)	FLOPs (G)
ResNet50	98.6	98.4	0.99	6.0	98	4.1
EfficientNet-B0	94.5	94.2	0.97	5.2	29	0.39
MobileNetV2	94.3	94.0	0.97	3.9	14	0.30
Hybrid CNN–LSTM	92.8	92.5	0.95	18.0	30	2.5
ViT-Small	95.4	95.1	0.98	12.3	44	11.6

ResNet50 [28] achieves peak static accuracy and ROC-AUC, but EfficientNet-B0 [29] and MobileNetV2 deliver sub-6 ms latency with ≈94% accuracy. CNN–LSTM hybrids excel in dynamic gesture recognition (~92.8%) but incur higher inference times.

Table 5 TSC Performance Summary

Model	Acc (%)	mAP@50 (%)	ROC-AUC	Latency (ms)	Size (MB)	FLOPs (G)
MicronNet	96.7	—	0.95	2.8	1.2	0.05
ShuffleNetV2	97.8	—	0.96	3.1	7.4	0.15
ResNet50	99.28	—	0.99	6.0	98	4.1
E-MobileViT	99.61	—	0.98	3.5	25	12.0
YOLOv5s	98.9	83.2	0.99	7.2	27	17.5

MicronNet's ultra-light design suits microcontroller deployment. E-MobileViT [30] balances top-tier accuracy with low-footprint inference. YOLO detectors offer combined classification and localization at the cost of higher computational load.

Table 6 Robustness Analysis

Noise Type	ResNet50 Δ (%)	MicronNet Δ (%)	E-MobileViT Δ (%)	ViT-Small Δ (%)	YOLOv5s Δ (%)
Gaussian (σ=0.1)	1.2	1.5	0.8	1.1	1.2
Salt-Pepper (5%)	8.5	6.2	7.0	5.4	4.8
Motion Blur	5.2	3.2	4.0	3.6	2.9
Shot Noise	2.7	2.8	3.0	2.5	2.1
Occlusion (10%)	7.8	9.5	5.2	6.8	7.4

Salt-and-pepper and occlusion severely degrade performance, especially for lightweight models. Incorporating targeted augmentations (CutMix, Random Erasing) is vital for real-world robustness.

4.1. Future Work

Edge Optimization: Implement mixed-precision training, 8-bit quantization, and structured pruning on ResNet50, E-MobileViT, and ViT-Small to target <2 ms inference on ARM CPUs while maintaining ≥95% of original accuracy.

Adaptive Augmentation: Develop sensor-aware augmentation pipelines that simulate realistic motion blur, lighting changes, and occlusions based on real-world deployment environments (e.g., nighttime driving, indoor signing under low light).

Multimodal Integration: Fuse RGB, depth, and inertial sensor data to enhance SLR in occluded and low-contrast scenarios, and integrate LiDAR-based depth cues for TSC under adverse weather.

Continual and Federated Learning: Design privacy-preserving federated frameworks to update SLR models from distributed user data, and continual learning schemes to incorporate new sign classes and regional traffic sign variants without catastrophic forgetting.

User-Centric Evaluation: Conduct longitudinal user studies with deaf participants to assess translation latency, accuracy, and user satisfaction; perform on-road trials for TSC in partnership with automotive OEMs to validate system reliability under diverse driving conditions.

Improving Robustness: Many models struggle with variations in lighting, hand shape, and user differences. Improving robustness is crucial. Techniques such as data augmentation, adversarial training, and domain adaptation could be explored. Data augmentation can generate synthetic data, while adversarial training makes models more resistant to noisy inputs. Domain adaptation helps models generalize better to new datasets.

Real-time Performance: Real-time performance is essential for practical applications. Optimizing models for speed and efficiency is key. This could involve exploring more efficient architectures (like MobileNet) or using model compression techniques.

Enhancing YOLOv5 Model: While our YOLOv5 model shows promising results, further improvements are possible. Techniques such as transfer learning, more sophisticated loss functions (like focal loss), or different backbone architectures could enhance performance. Further hyperparameter tuning and data augmentation tailored to the YOLOv5 architecture could also improve accuracy. The exploration of different anchor box sizes and aspect ratios within the YOLOv5 architecture could also be beneficial. Investigating different data augmentation strategies, such as MixUp or CutMix, could further enhance model robustness and generalization capabilities.

Limitations

- Dataset Bias: Sign Language MNIST and ASL Fingerspelling datasets lack diversity in signer demographics and environmental conditions, potentially limiting generalization to underrepresented skin tones and backgrounds.
- Controlled Perturbations: Noise injection experiments use synthetic perturbations that may not fully capture complex real-world artifacts such as rain droplets on camera lenses or motion parallax effects.
- Hardware Constraints: Latency and energy measurements are reported for specific ARM and GPU platforms; performance may vary significantly on alternative edge devices (e.g., microcontrollers, smartphones).
- Evaluation Scope: Cross-dataset transfer focuses on GTSRB→BelgiumTS and TT100K→GTSRB; additional regional sign datasets (e.g., Chinese, Indian traffic signs) are not evaluated due to resource constraints.
- Model Interpretability: Deep architectures lack explainability mechanisms, posing challenges for safety-critical applications; future work should explore attention visualization and saliency mapping for both SLR and TSC.

5. Conclusion

This study undertook a comprehensive examination of recent advancements in Sign Language Recognition (SLR) and Traffic Sign Classification (TSC), two domains that illustrate the practical and social potential of modern computer vision systems. By reviewing 30 state-of-the-art models and benchmarking their performance across multiple datasets, this report has brought to light the evolving capabilities—and limitations—of deep learning architectures in real-world, time-critical applications.

Across both domains, one clear trend emerges: accuracy alone is no longer a sufficient metric for success. Models must now balance recognition performance with factors such as computational cost, latency, and robustness to noise and variation. Our findings confirm that while high-capacity architectures like ResNet50 and transformer-based models deliver exceptional performance under ideal conditions, their complexity and resource requirements limit deployment on edge devices. Conversely, lightweight models such as MobileNetV2 and EfficientNet offer promising trade-offs, enabling deployment on mobile and embedded platforms with minimal accuracy loss.

In the context of SLR, models incorporating temporal dynamics—such as CNN-LSTM hybrids, transformer networks, and graph neural networks—proved particularly effective for capturing the fluid nature of sign sequences. The addition of depth sensors, motion tracking, and NLP-based translation modules further enhanced usability, enabling more natural and context-aware communication tools. Still, the field faces ongoing challenges, including the recognition of overlapping gestures, signer variability, and the lack of diverse, high-quality datasets.

For TSC, the emphasis on real-time performance and environmental variability brought lightweight models and efficient detectors like YOLOv5 to the forefront. These systems demonstrated strong accuracy under normal conditions but struggled under occlusion, motion blur, and low-contrast scenarios—conditions frequently encountered on real roads. This highlights the importance of domain-specific augmentation, robust training pipelines, and cross-dataset evaluation for improving system reliability.

Importantly, the study also underscored the significance of practical considerations often overlooked in academic benchmarks. Metrics such as model size, power consumption, and inference speed directly impact the viability of real-world deployment. The need for transparent evaluation across these dimensions is critical if research is to translate into inclusive assistive tools and safe autonomous systems.

Looking ahead, it is clear that the next wave of innovation must focus on integration and adaptability—combining visual recognition with language understanding, adapting to user-specific variations, and continuously learning from new inputs without retraining from scratch. Equally, models must become more explainable, especially in high-stakes contexts like road safety and human communication, where transparency and trust are essential.

In closing, while deep learning has brought both SLR and TSC closer to practical deployment, bridging the remaining gaps will require continued interdisciplinary effort—drawing from computer vision, human-computer interaction, embedded systems, and social accessibility research. With thoughtful development and user-centered design, these technologies hold the promise to not only automate but also humanize machine perception in meaningful ways.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed..

References

[1] S. Kumar, A. Walia, and K. Singh, "DeepSign: Sign language detection and recognition using deep learning," Electronics, vol. 10, no. 11, p. 1780, 2021.

[2] I. Ahmad, X. Wang, R. Zhu, and H. Zhang, "Sign language recognition using modified deep learning network and hybrid optimization: A hybrid optimizer (HO) based optimized CNNs-LSTM approach," Scientific Reports, vol. 13, p. 15537, 2023.

[3] B. Garcia and S. A. Viesca, "Real-time American sign language recognition with convolutional neural networks," Convolutional Neural Networks for Visual Recognition, vol. 2, pp. 225–232, 2016.

[4] R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in Proc. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition, 1998, pp. 558–567.

[5] J. Zhang, Q. Wang, Q. Wang, and Z. Zheng, "Multimodal fusion framework based on statistical attention and contrastive attention for sign language recognition," IEEE Trans. Mobile Comput., vol. 23, no. 2, pp. 1431–1443, 2023.

[6] H. Brashear, T. Starner, P. Lukowicz, and H. Junker, "Using multiple sensors for mobile sign language recognition," in Proc. 7th IEEE Int. Symp. Wearable Computers, 2003, p. 45.

[7] N. S. Khan, A. Abid, and K. Abid, "A novel natural language processing (NLP)–based machine translation model for English to Pakistan sign language translation," Cogn. Comput., vol. 12, pp. 748–765, 2020.

[8] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, "Wearable sensor-based sign language recognition: A comprehensive review," IEEE Rev. Biomed. Eng., vol. 14, pp. 82–97, 2020.

[9] Q. De Smedt, H. Wannous, and J. P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2016, pp. 1–9.

[10] M. O. Khairandish et al., "A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images," IRBM, vol. 43, no. 4, pp. 290–299, 2022.

[11] S. Masood, A. Srivastava, H. C. Thuwal, and M. Ahmad, "Real-time sign language gesture (word) recognition from video sequences using CNN and RNN," in Intelligent Engineering Informatics, Singapore: Springer, 2018, pp. 623–632.

[12] B. Batte, "AI-powered sign language translation system: A deep learning approach to enhancing inclusive communication and accessibility in low-resource contexts," SSRN, 2025. [Online]. Available: SSRN

[13] X. Chen et al., "Hand gesture recognition based on surface electromyography using convolutional neural network with transfer learning method," IEEE J. Biomed. Health Inform., vol. 25, no. 4, pp. 1292–1304, 2020.

[14] M. Parelli et al., "Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos," in Computer Vision–ECCV 2020 Workshops, Springer, 2020, pp. 249–263.

[15] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in Proc. IEEE RO-MAN, 2012, pp. 411–417.

[16] J. Shin et al., "Korean sign language recognition using transformer-based deep neural network," Appl. Sci., vol. 13, no. 5, p. 3029, 2023.

[17] A. Baihan et al., “Sign language recognition using modified deep learning network and hybrid optimization: a hybrid optimizer (HO) based optimized CNNSa-LSTM approach,” Scientific Reports, vol. 14, no. 1, p. 26111, 2024.

[18] M. Alaftekin, I. Pacal, and K. Cicek, “Real-time sign language recognition based on YOLO algorithm,” Neural Comput. Appl., vol. 36, no. 14, pp. 7609–7624, 2024.

[19] J. Kan et al., “Sign language translation with hierarchical spatio-temporal graph neural network,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2022, pp. 3367–3376.

[20] A. Musa et al., “Lightweight Deep Learning Models For Edge Devices—A Survey,” Int. J. Comput. Inf. Syst. Ind. Manag. Appl., vol. 17, pp. 18–18, 2025.

[21] P. V. V. Kishore et al., “Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks,” in Proc. IEEE 6th Int. Conf. Adv. Comput. (IACC), 2016, pp. 346–351.

[22] A. Farhadi, D. Forsyth, and R. White, “Transfer learning in sign language,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.

[23] M. M. Hassan, S. Ullah, M. S. Hossain, and A. Alamri, “Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism,” Electronics, vol. 13, no. 7, p. 1229, 2024.

[24] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “The German traffic sign recognition benchmark: A multi-class classification competition,” in Proc. Int. Joint Conf. Neural Netw., 2011, pp. 1453–1460.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[27] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint, arXiv:1409.1556, 2014.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[29] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 6105–6114.

[30] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint, arXiv:2010.11929, 2020.