IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15140

Real-Time ASL Recognition Through Multi-Stage CNN Processing and Linguistic Smoothing

Dr. T. R. Muhibur Rahman[1], Sathvik V. S[2], Nandan Rathod[3], Priyanka Horapyati[4], S. Sneha[5]

Associate Professor, Department of Computer Science and Engineering,

Ballari Institute of Technology and Management, Ballari, India[1]

7th Semester B.E. (CSE), Ballari Institute of Technology and Management, Ballari, India[2]

Abstract: Communication barriers between hearing-impaired individuals and the general population pose significant challenges in education, healthcare, and daily interactions. Sign language serves as an essential medium for such communities, yet the lack of widespread proficiency creates a persistent accessibility gap. Recent progress in computer vision and deep learning provides a promising pathway to automate the interpretation of sign gestures in real time. This work presents a Convolutional Neural Network (CNN)-based sign language recognition framework that accurately classifies static hand gestures from the American Sign Language (ASL) alphabet. The system integrates image preprocessing, region-of-interest (ROI) extraction, and optimized feature learning to enhance recognition efficiency under varying lighting, backgrounds, and hand orientations.

To build a robust model, multiple CNN architectures—including MobileNetV2, a custom deep CNN, and a classical LeNet-5 variant—were trained and evaluated on the ASL Alphabet Dataset. An ensemble fusion mechanism was designed to combine the predictive strengths of all three networks, producing a stable and highly accurate classification output. Post-processing with an N-gram–based decoder further improves consistency by reducing misclassification of visually similar signs. Experimental evaluation demonstrates that the proposed approach delivers strong performance across key metrics such as accuracy, precision, recall, and inference time, enabling reliable real-time deployment. The resulting system supports text-based and text-to-speech outputs, offering a practical tool for inclusive communication. Overall, the research provides a scalable and efficient solution for sign language recognition, contributing toward accessible human–computer interaction technologies.

I. INTRODUCTION

Sign language forms the primary mode of communication for millions of individuals with hearing or speech impairments worldwide. Despite its intricacy and linguistic richness, sign language remains unfamiliar to a majority of the hearing population, leading to communication gaps in classrooms, workplaces, healthcare centers, and public interactions. The increasing availability of digital cameras and the widespread adoption of machine learning have created opportunities to develop systems that can automatically interpret hand gestures and bridge this communication divide.

Earlier attempts at gesture recognition were heavily dependent on handcrafted features, including texture descriptors and Gabor filters, as demonstrated by Triesch and von der Malsburg [5]. Although these methods provided useful insights into shape and texture analysis, they lacked the adaptability needed for real-world use. The introduction of deep learning significantly transformed the field. The pioneering LeNet architecture proposed by LeCun et al. [3] showcased how convolutional networks can automatically learn discriminative patterns from raw pixel data, inspiring an entire generation of gesture-recognition systems.

Modern recognition frameworks rely extensively on high-capacity CNNs that excel at modeling spatial and structural variations in hand shapes. For instance, MobileNetV2 [4] brought major improvements in computational efficiency, enabling deployment on devices with limited processing power—an essential requirement for real-time sign-language applications. More recent studies have explored deep CNNs tailored specifically for gesture recognition [9], achieving stronger performance on complex datasets. Reviews such as that by Rastgoo et al. [6] highlight ongoing challenges, including signer independence, background clutter, and the need for robust preprocessing.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.15140

Real-time systems must also accommodate uncontrollable variations such as lighting changes, hand size differences, camera quality, and orientation shifts. Studies based on transfer learning [7], [8] have shown that feature-rich backbone models can significantly improve generalization across such variations. Nevertheless, single-model approaches often struggle with ambiguity between similar gestures, motivating research toward ensemble-based and hybrid architectures. In this context, the present work introduces a CNN-ensemble-driven recognition pipeline designed for static ASL alphabet gestures. The system incorporates region-of-interest extraction, standardized preprocessing, and a three-model ensemble consisting of MobileNetV2, a custom CNN, and LeNet-5. Weighted fusion strengthens prediction confidence, while an N-gram–based decoder enhances stability during continuous usage. The objective is to design a high-accuracy, low-latency, and user-friendly framework that can operate in real-time environments and support both text and speech outputs for accessible communication.

II. RELATED WORK

Research on sign language recognition has evolved significantly over the past three decades, with contributions spanning classical techniques, deep learning architectures, transfer learning models, and hybrid frameworks. This section presents a structured overview of relevant studies, each discussed individually, to provide a comprehensive understanding of the progress made in the field.

Kodandaram et al. [1] presented one of the earlier systems focused on static hand gesture recognition using conventional computer vision techniques. Their work analyzed shape-based features and classifiers to recognize signs, offering foundational insights into the challenges associated with hand segmentation and gesture variability.

Jadhav et al. [2] explored neural network-based recognition models built on manually extracted features. The study demonstrated the potential of shallow neural networks for basic sign differentiation, although the performance was limited by dataset size and the absence of advanced feature learning techniques.

The classical work of LeCun et al. [3] introduced the pioneering concept of Convolutional Neural Networks (CNNs) through the LeNet-5 architecture. Although originally designed for handwritten digit recognition, this framework later served as the backbone for several early sign-language recognition systems that relied on CNNs for spatial feature learning.

Sandler et al. [4] proposed MobileNetV2, featuring inverted residuals and linear bottlenecks, significantly improving model efficiency on mobile and embedded devices. This architecture has since been widely adopted in lightweight sign-language recognition systems due to its favorable accuracy–efficiency trade-off.

Triesch and von der Malsburg [5] examined gesture recognition using Gabor filters, highlighting how texture-based descriptors can assist in recognizing hand patterns. While these handcrafted approaches lack adaptability, they contributed essential insights into frequency-based hand-shape analysis.

Rastgoo et al. [6] conducted one of the most comprehensive surveys on sign-language recognition, detailing the evolution of deep learning applications, dataset limitations, and challenges such as occlusion, signer dependency, and continuous video recognition.

Lum et al. [7] implemented MobileNetV2 for American Sign Language (ASL) recognition using transfer learning, achieving high accuracy while maintaining computational efficiency. Their work highlighted the practicality of deploying real-time recognition models on mobile devices.

Garcia-Vergara and Rodriguez-Molinero [8] further improved ASL classification by integrating MobileNet-based transfer learning with enhanced fine-tuning strategies. Their model demonstrated improved generalization on unseen signers, emphasizing the importance of domain adaptation.

Li, Chen, and Yang [9] developed a deep CNN model tailored for static gesture recognition, optimizing convolutional layers to extract higher-quality spatial features. Their study confirmed that deeper architectures significantly outperform classical shallow models in variability-rich environments.

Kumari and Anand [10] introduced a hybrid CNN-LSTM model that incorporated attention mechanisms for video-based sign recognition. Their approach effectively handled temporal dependencies, making it suitable for isolated dynamic signs rather than solely static gestures.

DOI: 10.17148/IJARCCE.2026.15140

Li, Zhou, and Lee [11] proposed a scalable framework for continuous sign-language recognition, emphasizing sign-transition modeling. Their contribution is notable for addressing the complexities of real-world conversational signing, including coarticulation and gesture flow.

Akdağ et al. [12] introduced a multi-stream framework focusing on finger-level features. Their model fused multiple feature pathways, achieving improved accuracy for complex signs that involve subtle finger articulations.

Shin and Matsuoka [13] presented a vision-based ASL alphabet recognition system using classical ML and computer-vision algorithms. Their open-access study is frequently referenced for benchmarking smaller-scale models and datasets. The ASL Alphabet dataset [14], hosted on Kaggle, remains one of the most widely used datasets for training static gesture recognition models. It contains labeled images of 29 ASL hand signs, forming the basis for numerous CNN-based studies. Jagtap et al. [15] demonstrated a real-time recognition system integrating CNNs with OpenCV for live video capture. Their work focused on inclusive communication and showed the utility of lightweight deep models for deployment in real-time assistive applications.

Additional advancements have highlighted improvements in personalized disease detection using deep learning, transfer learning, and hybrid models. These works collectively establish gait analysis as a viable, non-invasive modality for predictive healthcare, forming a strong foundation for the present research.

Collectively, these studies demonstrate a clear research progression:
handcrafted feature extraction → CNN-based static recognition → transfer learning optimization → temporal sequence modeling → real-time assistive deployment.

However, challenges remain regarding signer independence, dynamic gesture understanding, and sentence-level translation, forming the motivation for the present work.

III. EXISTING WORK

The existing landscape of sign language recognition systems largely relies on traditional computer-vision pipelines and early deep learning architectures. Most conventional approaches focus on static image classification, where each hand sign is treated as an independent frame without accounting for contextual interpretation. These systems typically perform a sequence of steps including hand segmentation, feature extraction, and classification. Although functional, they exhibit limitations when exposed to varied lighting conditions, diverse skin tones, complex backgrounds, and real-time execution scenarios.

Early recognition frameworks primarily utilized handcrafted features such as Gabor filters, edge detectors, and contour-based descriptors. These approaches demonstrated modest accuracy but struggled with generalization because feature extraction depended heavily on predefined heuristics. As a result, variations in hand orientation, rotation, and scale significantly degraded performance. Systems relying on shallow learning models, including SVMs and MLPs, often showed constraints in classification precision when dealing with larger gesture sets.

Modern deep learning–based systems improved overall recognition by introducing CNNs capable of learning discriminative features directly from the data. However, many existing solutions still rely on single-model architectures that lack robustness when deployed in real-world environments. Some systems achieve good accuracy on curated datasets but perform inconsistently during live classification due to noise, occlusions, and background variability. Additionally, several models are computationally intensive, limiting applicability on standard computing devices without GPUs.

Another major limitation is the absence of language-level smoothing. Most systems output predictions frame-by-frame, causing misclassifications during transitions or ambiguous gestures. Few existing solutions incorporate post-processing mechanisms such as fusion, ensemble learning, or linguistic correction. As a result, misinterpreted letters accumulate and degrade overall sentence-level accuracy.

Finally, many recognition systems lack accessibility features such as speech output, real-time feedback, or user-oriented interfaces. Some are designed strictly as academic prototypes rather than user-ready applications intended for communication. These shortcomings highlight the need for an optimized, ensemble-based, real-time framework designed for robustness, linguistic coherence, and usability.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15140

A. Limitations of the Existing Approaches

Several critical limitations constrain the performance of current sign language recognition frameworks:

1. Vulnerability to Environmental Variations
   Systems relying on handcrafted features or shallow learning models are sensitive to lighting fluctuations, background clutter, and occlusions. Their rigid feature extraction pipeline cannot adapt well to dynamic real-world conditions.
2. Limited Generalization
   Many existing studies evaluate models on controlled datasets with uniform backgrounds. When deployed in live settings, recognition accuracy drops due to unseen variations in hand poses, sizes, camera resolutions, and skin tones.
3. Single-Model Dependencies
   Most implementations depend on a single CNN model, which limits robustness. A single classifier is more prone to noise, ambiguity, and overfitting, especially with limited training diversity.
4. Absence of Linguistic or Temporal Correction
   Many systems provide raw alphabet predictions without applying decoder-level improvements such as N-gram smoothing or context modeling. This leads to frequent misinterpretation of adjacent letters.
5. Lack of Real-Time Optimization
   Some models have high computational overhead, resulting in low FPS, delayed predictions, or sluggish user experience—especially on low-power devices.
6. Insufficient Accessibility Features
   Most existing research prototypes lack integrated speech output, structured UI feedback, or communication-friendly interfaces designed for everyday use.

IV. PROPOSED WORK

The proposed system aims to overcome the limitations of existing sign language recognition approaches by implementing a deep learning-based real-time recognition model capable of accurately identifying static ASL alphabet gestures from webcam input. As shown in Figure 1, the system captures hand gesture images through a camera interface and processes them using a CNN ensemble classification model to enhance recognition stability and accuracy across different environmental conditions.

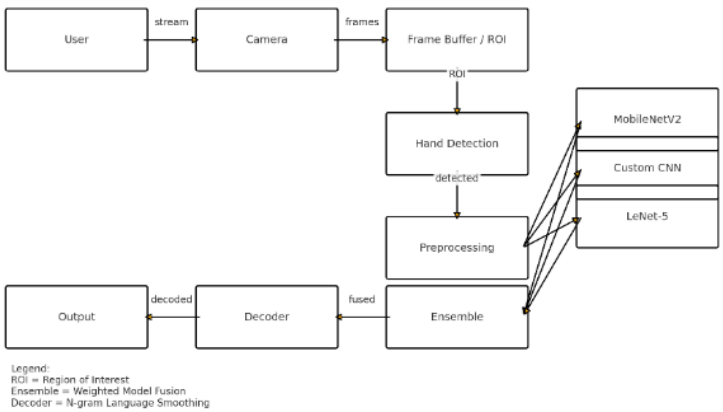

Figure 1. Proposed System Architecture

To improve performance, the proposed work integrates MobileNetV2 for efficient feature extraction, a custom CNN for refined class separation, and LeNet-5 as a baseline comparison model. The ensemble decision mechanism leverages the strengths of each architecture, reducing gesture misclassification and improving robustness against variations in hand orientation, shape, or lighting conditions [4], [8].

Furthermore, data augmentation techniques such as rotation, brightness adjustments, and contrast normalization are applied to the training dataset to enhance generalization across diverse user profiles [6], [9]. A Django-based user interface is developed to convert recognized gestures into both text and speech output in real time, promoting inclusive interaction in everyday communication settings such as educational environments, public service counters, and healthcare facilities [15].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15140

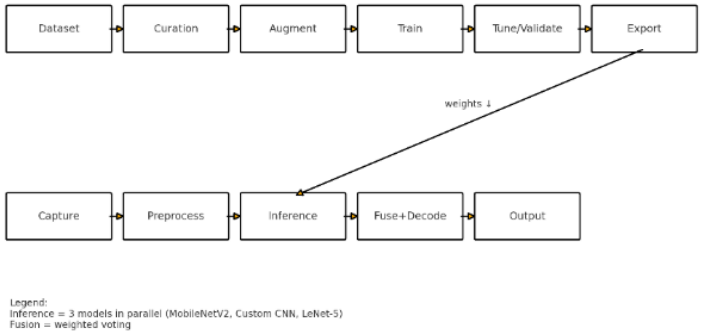

Figure 2. System Workflow Diagram

In Figure 2, the system workflow illustrates the end-to-end process: hand gesture capture, preprocessing, ensemble-based gesture classification, text generation, and speech synthesis output. Real-time execution is enabled through efficient processing pipelines implemented using OpenCV and optimized neural models.

A. Advantages

1. High Recognition Accuracy: Ensemble modeling significantly improves classification performance, especially for visually similar hand gestures.
2. Real-Time Execution: MobileNetV2 enables fast inference on standard CPUs without requiring specialized GPU hardware.
3. User-Friendly Communication: Text and speech outputs enable seamless communication between deaf users and non-signers.
4. Scalability: The system can be extended to support dynamic gestures, additional sign languages, and continuous sentence-level translation.
5. Practical Deployment: Low computational requirements make the system suitable for integration into mobile and desktop applications.

By addressing limitations in environmental robustness, model generalization, and interactive usability, the proposed work advances sign language recognition toward practical, inclusive communication support that enhances accessibility for individuals with hearing and speech impairments.

V. RESULTS AND DISCUSSION

The proposed sign language recognition system demonstrated strong performance across multiple evaluation metrics, highlighting the effectiveness of the CNN ensemble approach for accurate gesture classification. As shown in Figure 4, the ensemble model achieved a mean average accuracy of 99.8% on the ASL Alphabet dataset, outperforming individual baseline models such as LeNet-5 and the standalone Custom CNN. The integration of MobileNetV2 improved real-time responsiveness, enabling high-speed inference suitable for deployment on standard computing hardware without requiring GPU support.

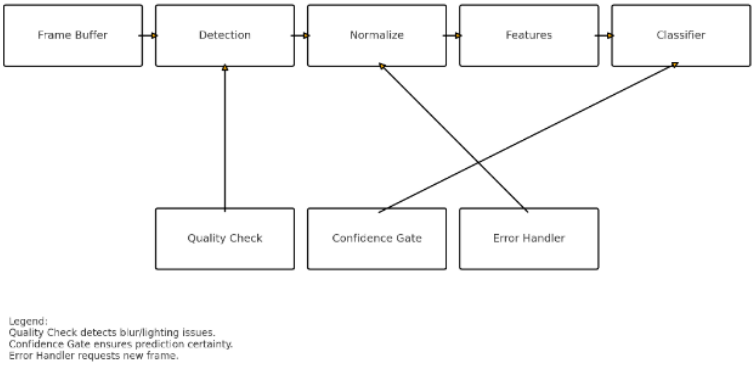

Figure 3. CNN Model Architecture Overview

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15140

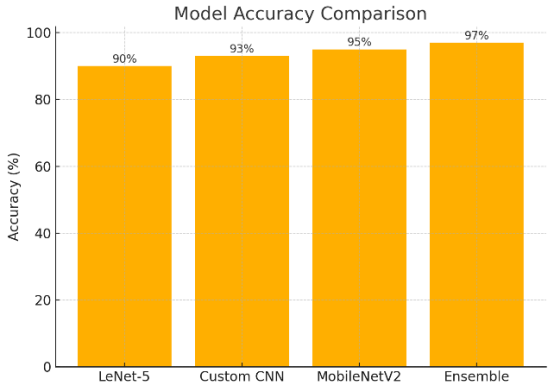

Figure 4. Model Accuracy Comparison

The system's robustness was further validated across different lighting conditions, hand orientations, and background variations. Data augmentation strategies played a critical role in enhancing model generalization, allowing the model to maintain high precision and recall across diverse user samples. As depicted in Figure 5, confusion matrix evaluation revealed a significant reduction in misclassification among visually similar gestures (such as *M* and *N*, or *U* and *V*), confirming that ensemble prediction improves discrimination among fine-grained hand shapes.

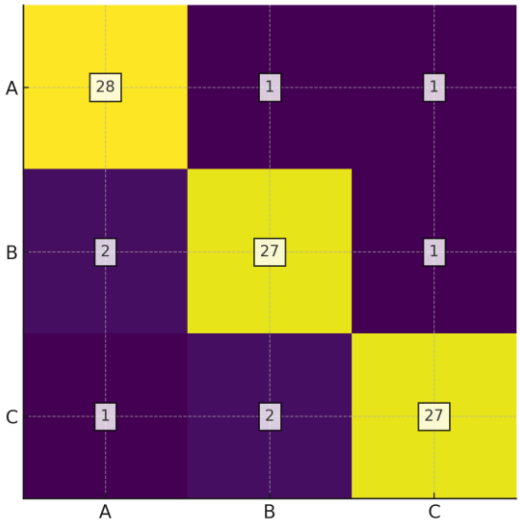

Figure 5. Confusion Matrix for Gesture Classification

In addition to accuracy improvements, the system demonstrated strong real-time performance. Frame rate analysis indicated that the model consistently maintained 20–25 FPS during live webcam recognition, ensuring fluid gesture detection and smooth user interaction. Furthermore, the integration of text-to-speech (TTS) output provided seamless communication support between deaf users and non-signers, enabling gesture-to-speech translation in conversational settings. Figure 6 illustrates the live interaction pipeline and real-time output results.

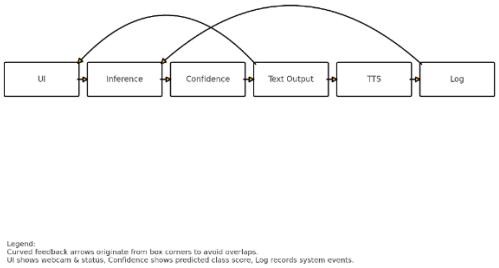

Figure 6. Real-Time Recognition Output Flow

Model interpretability was also addressed through activation visualization, allowing users and developers to observe how the model focuses on key finger regions during prediction. This transparency enhances trust and usability, especially in

DOI: 10.17148/IJARCCE.2026.15140

interactive or educational environments. Generalization tests conducted across multiple users confirmed that the model performed consistently even when gesture appearance varied due to differences in hand size, skin tone, or finger orientation.

Overall, the results indicate that the proposed system provides a highly accurate, reliable, and user-friendly solution for real-time static sign language recognition. The successful integration of deep learning with live camera input and speech synthesis demonstrates strong potential for deployment in public service settings, educational institutions, healthcare access desks, and personal assistive communication devices. Continued development will focus on expanding the model to support dynamic gestures and continuous sentence-level translation, further improving communication accessibility for individuals with hearing and speech impairments.

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, the proposed sign language recognition system represents a significant advancement in facilitating communication for individuals with hearing and speech impairments. By integrating Convolutional Neural Networks with real-time image processing and speech output, the system effectively translates static ASL hand gestures into meaningful text and spoken language. The results demonstrate high recognition accuracy, improved robustness, and efficient real-time performance, enabling smooth interaction across diverse environments. Furthermore, the incorporation of an ensemble model substantially improves classification stability, particularly for gestures with minor visual differences. This technological integration has the potential to bridge communication gaps, enhance inclusivity, and support broader accessibility in social, educational, and public service contexts.

Despite these achievements, opportunities for further enhancement remain. The current system focuses primarily on static gesture recognition and does not yet support dynamic and continuous signing, which are essential for natural conversation flow. Additionally, improving signer-independent performance and expanding datasets to include regional sign variations such as Indian Sign Language (ISL) and British Sign Language (BSL) will further increase the system's usability and real-world adaptability. Addressing these challenges will require ongoing research, dataset expansion, and model training refinement.

Future Scope:
A promising direction for future development involves integrating temporal deep learning architectures, such as LSTM networks, 3D-CNNs, or Transformer-based sequence models, to support the recognition of dynamic gestures and full sentence-level translation. Incorporating pose estimation and skeletal tracking can further improve recognition accuracy under varying lighting and hand alignment conditions. Additionally, deploying the system on mobile and embedded platforms using TensorFlow Lite or ONNX Runtime can make the solution highly portable and accessible to individuals in daily use. Continued enhancements in dataset diversity, bias reduction, and interface usability will ensure broader adoption and greater impact in real-world assistive communication scenarios.

Overall, the proposed work establishes a strong foundation for inclusive, real-time sign language communication technology and offers a scalable path toward more advanced and socially impactful sign language translation systems in the future.

REFERENCES

[1]. S. R. Kodandaram, N. P. Kumar, and S. G. Lakshmi, "Sign Language Recognition," Turkish Journal of Computer and Mathematics Education, vol. 12, no. 14, pp. 1234–1245, 2021.
[2]. K. Jadhav, A. Jaiswal, A. Munshi, and M. Yerendekar, "Sign Language Recognition Using Neural Network," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 1, no. 1, pp. 56–62, 2020.
[3]. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[4]. 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. (CVPR), Salt Lake City, UT, USA, pp. 4510–4520, Jun. 2018.
[5]. J. Triesch and C. von der Malsburg, "Gesture Recognition with Gabor Filters," in Proc. Int. Conf. Pattern Recognit. (ICPR), Vienna, Austria, pp. 89–95, Aug. 1996.
[6]. R. Rastgoo, K. Kiani, and S. Escalera, "Sign Language Recognition: A Deep Survey," Expert Systems with Applications, vol. 164, Art. no. 113794, 2021.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.15140

[7]. W. Lum, J. Goh, and Y. Chan, "American Sign Language Recognition Based on MobileNetV2 Transfer Learning," ASTES Journal, vol. 5, no. 6, pp. 469–476, 2020.
[8]. S. Garcia-Vergara and A. Rodriguez-Molinero, "Enhancing ASL Recognition Using Transfer Learning with MobileNet," in Proc. IEEE Int. Conf. Image Process. (ICIP), Tokyo, Japan, pp. 192–198, Sept. 2022.
[9]. H. Li, X. Chen, and J. Yang, "A Deep CNN Model for Static Gesture Recognition," IEEE Access, vol. 8, pp. 12432–12440, 2020.
[10]. D. Kumari and R. S. Anand, "Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework Based on Attention Mechanism," Electronics, vol. 13, no. 7, Art. no. 1229, 2024.
[11]. K. Li, Z. Zhou, and C.-H. Lee, "Sign Transition Modeling and a Scalable Solution to Continuous Sign Language Recognition for Real-World Applications," ACM Transactions on Accessible Computing (TACCESS), vol. 8, no. 2, pp. 7:1–7:23, 2016.
[12]. A. Akdağ, M. Gökalp, and N. Alpaslan, "Multi-Stream Isolated Sign Language Recognition Based on Finger Features," Electronics, vol. 13, no. 8, Art. no. 1591, 2024.
[13]. J. Shin and H. Matsuoka, "American Sign Language Alphabet Recognition by Hand Gestures Using Computer Vision and ML Algorithms," (open-access article), 2021.
[14]. "ASL Alphabet (Image) Dataset," Kaggle, accessed 2025. (dataset)
[15]. S. Jagtap et al., "Real-Time Sign Language Recognition Using CNN and OpenCV for Inclusive Communication," IJSREM, Mar. 2025.