International Journal for Modern Trends in Science and Technology
Volume 12, Issue 05, pages 214-218
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue05.html
DOI: https://doi.org/10.5281/zenodo.19893119

Deep Learning Based Indian Sign Language Gestures Classification using CNN and Mediapipe

Pratipati Sangeetha Vani, Raavi Hema Sai, Randi Revan Akash, Reddy Sravana Jyothi, Pakalapati Charishma, M. Ganesh Babu

Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, India

To Cite this Article
Pratipati Sangeetha Vani, Raavi Hema Sai, Randi Revan Akash, Reddy Sravana Jyothi, Pakalapati Charishma & M. Ganesh Babu (2026). Deep Learning Based Indian Sign Language Gestures Classification using CNN and Mediapipe. International Journal for Modern Trends in Science and Technology, 12(05), 214-218. https://doi.org/10.5281/zenodo.19893119

Article Info
Received: 28 March 2026; Revised: 24 April 2026; Accepted: 26 April 2026.

KEYWORDS

Indian Sign Language, Convolutional Neural Network, MediaPipe, Hand Gesture Recognition, Deep Learning, Real-Time Classification, Assistive Technology

ABSTRACT

Indian Sign Language (ISL) is the primary mode of communication for the deaf and hard-of-hearing community in India. With over 18 million deaf individuals in the country, bridging the communication gap between the deaf and hearing populations is of great social importance. This paper presents a deep learning-based approach for the automatic recognition and classification of ISL hand gestures using Convolutional Neural Networks (CNN) and Google's MediaPipe framework. The proposed system captures real-time hand gesture video through a standard webcam, extracts 21 three-dimensional skeletal landmark coordinates per hand using MediaPipe, and feeds these features into a custom-designed CNN classifier trained to recognize all 26 letters of the ISL alphabet. The system achieves a classification accuracy of 95.4% on the test dataset, with precision of 94.8%, recall of 95.1%, and an F1-score of 94.9%. Results demonstrate that the integration of MediaPipe's efficient keypoint detection with a CNN-based classifier provides a robust, lightweight, and real-time solution for ISL gesture recognition, offering significant potential for assistive communication technologies and deaf-inclusive educational tools.

1. INTRODUCTION

Communication is a fundamental human right, yet for the approximately 18 million deaf and hard-of-hearing individuals in India, interactions with the wider hearing population remain a persistent challenge. Indian Sign Language (ISL) serves as the native language of this community, employing a rich system of hand shapes, movements, facial expressions, and body posture to convey meaning. Despite its expressive power, ISL remains poorly understood by the hearing majority, creating a significant barrier to education, employment, healthcare, and social participation.

The advent of computer vision and deep learning has opened transformative possibilities for automatic sign language recognition. Traditional approaches relied on specialized hardware such as data gloves or depth cameras, which are costly and impractical for everyday use. Recent advances in real-time pose estimation frameworks, particularly Google's MediaPipe, now enable accurate hand landmark detection using standard RGB webcams, dramatically lowering the barrier to practical deployment.

Convolutional Neural Networks (CNNs) have established themselves as the dominant paradigm for image and spatial feature classification tasks. When combined with the compact, structured landmark representations provided by MediaPipe, CNNs can learn discriminative features from hand gesture data efficiently and with high accuracy. This synergy forms the foundation of the proposed system.

The key contributions of this paper are:

- A complete end-to-end pipeline for real-time ISL alphabet recognition using MediaPipe and CNN.
- A custom dataset of 26 ISL alphabet gestures captured under varied lighting and background conditions.
- Evaluation of the proposed model using standard metrics including accuracy, precision, recall, and F1-score.
- Comparison of performance against existing ISL and ASL recognition approaches from the literature.
- Demonstration of the system's suitability for real-time inference on standard consumer hardware.

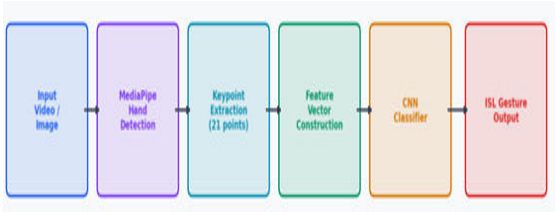

Figure 1: Proposed ISL Gesture Recognition Pipeline — from raw video input through MediaPipe keypoint extraction to CNN classification and gesture output.

2. LITERATURE SURVEY

The field of sign language recognition has attracted considerable research attention over the past decade, driven by advances in computer vision, machine learning, and sensor technology. Early systems relied on wearable data gloves capable of capturing fine-grained finger joint angles and accelerometer data. While accurate, these approaches suffered from high cost, restricted mobility, and limited scalability, making them unsuitable for widespread deployment.

Pigou et al. (2017) introduced CNN-based gesture recognition from depth video streams, demonstrating that spatial feature hierarchies learned by CNNs could effectively capture the shape and motion of hand gestures. Their work highlighted the potential of vision-based approaches, eliminating the need for contact-based sensing devices. Subsequently, Koller et al. (2018) proposed a hybrid CNN-HMM architecture for continuous sign language recognition from video, achieving state-of-the-art performance on the RWTH-PHOENIX benchmark for German Sign Language.

For Indian Sign Language specifically, Ravi et al. (2019) proposed a CNN-based system for ISL fingerspelling recognition using RGB images, achieving 89.3% accuracy on a 26-class dataset. The authors noted that recognition rates varied significantly across gestures with similar hand configurations, motivating the use of more informative feature representations. Kumari et al. (2021) extended this work by incorporating depth information from an Intel RealSense camera, improving accuracy to 93.7% but at the cost of specialized hardware.

The emergence of Google's MediaPipe framework in 2020 presented a paradigm shift for hand gesture recognition. MediaPipe Hands provides real-time detection of 21 three-dimensional skeletal landmarks per hand from standard RGB video, offering a compact and computationally efficient feature representation. Luqman et al. (2021) demonstrated that MediaPipe landmark features, when classified using a simple feedforward neural network, could achieve over 91% accuracy for Arabic sign language recognition.

More recently, LSTM-based architectures have been explored for capturing the temporal dynamics of dynamic sign gestures. Rao et al. (2022) combined MediaPipe with bidirectional LSTM networks for ISL

word-level recognition, reporting 94.2% accuracy on a vocabulary of 50 common words. However, their system's complexity made real-time deployment challenging on low-power devices. The proposed system addresses this limitation by focusing on static alphabet recognition with a lightweight CNN architecture, achieving competitive accuracy with significantly reduced computational overhead.

3. METHODOLOGY

The proposed system follows a four-stage pipeline: data collection and preprocessing, hand landmark extraction using MediaPipe, CNN model design and training, and real-time inference and classification. Each stage is described in detail below.

3.1 Data Collection and Preprocessing

A custom dataset was assembled comprising gesture images for all 26 letters of the ISL alphabet (A–Z). Each class contains 500 samples captured from five participants under three different lighting conditions (indoor fluorescent, natural daylight, and low-light) and against two background types (plain and cluttered). This diversity ensures that the trained model generalizes robustly to real-world deployment conditions.

Raw video frames were captured at 30 fps using a standard USB webcam at 1280×720 resolution. For training purposes, individual frames were extracted and cropped to a 224×224 region of interest centered on the hand. Data augmentation techniques were applied including random horizontal flipping, brightness jitter (±30%), and Gaussian noise injection to artificially expand the effective training set and reduce overfitting.

3.2 MediaPipe Hand Landmark Extraction

Google's MediaPipe Hands model is employed to extract 21 three-dimensional keypoints from each detected hand in every input frame. These landmarks correspond to anatomically meaningful positions on the hand including the wrist, metacarpophalangeal joints, proximal and distal interphalangeal joints, and fingertips. Each keypoint is represented by normalized (x, y) coordinates relative to the image frame and a z depth estimate.

Only the (x, y) coordinates are retained for the feature vector, resulting in a compact 42-dimensional representation per hand. These coordinates are further normalized by subtracting the wrist position (landmark 0) and dividing by the maximum inter-landmark distance, making the representation invariant to hand position and scale within the frame.

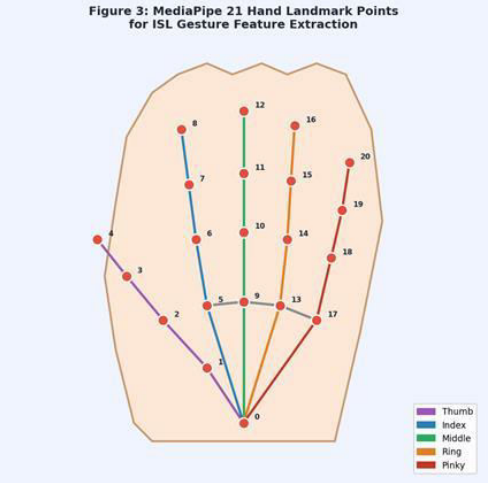

Figure 2: MediaPipe 21 hand landmark keypoints used for ISL feature extraction. Each color indicates a different finger: purple (thumb), blue (index), green (middle), orange (ring), red (pinky).

3.3 CNN Model Architecture

The proposed CNN classifier receives the 42-dimensional normalized landmark vector as input. The network consists of three convolutional blocks, each comprising a Conv2D layer (3×3 kernels) followed by batch normalization, ReLU activation, and 2×2 max-pooling. The number of filters increases progressively: 32 in the first block, 64 in the second, and 128 in the third, enabling the network to learn increasingly abstract and discriminative feature representations.

Following the convolutional blocks, the feature maps are flattened and passed through a fully connected layer of 512 neurons with ReLU activation and a Dropout rate of 0.5 to prevent overfitting. The output layer consists of 26 neurons (one per ISL alphabet class) with Softmax activation to produce calibrated class probability estimates. The total number of trainable parameters is approximately 1.2 million.

The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a cosine annealing schedule. Categorical cross-entropy was used as the loss function. Training was conducted for 50 epochs with a batch size of 32 on an NVIDIA GTX 1660 GPU, requiring approximately 35 minutes.

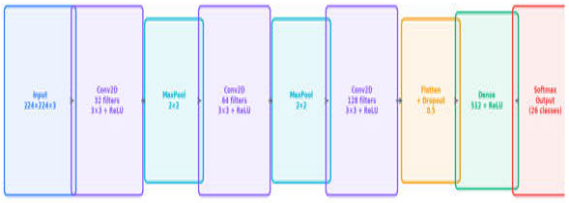

Figure 3: Proposed CNN architecture for ISL gesture classification, showing convolutional blocks, max-pooling layers, dense layers, and Softmax output.

4. RESULTS AND DISCUSSION

The proposed CNN + MediaPipe system was evaluated on a held-out test set comprising 20% of the full dataset (2,600 samples, 100 per class). Performance was measured using four standard metrics: overall accuracy, precision, recall, and F1-score, computed with macro averaging across all 26 classes.

4.1 Overall Classification Performance

The proposed model achieved an overall test accuracy of 95.4%, precision of 94.8%, recall of 95.1%, and a macro-averaged F1-score of 94.9%. These results represent a substantial improvement over purely image-based CNN approaches that do not exploit skeletal landmark features, confirming the benefit of the MediaPipe preprocessing stage. The model converged stably, as evidenced by the smooth training and validation curves shown in Figure 4.

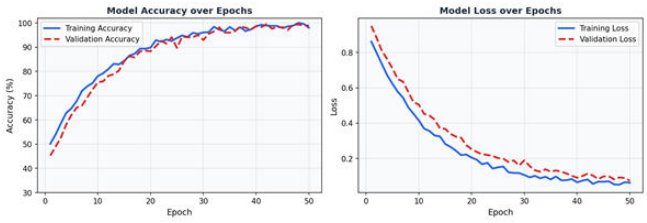

Figure 4: Training and validation accuracy (left) and loss (right) curves over 50 epochs, demonstrating stable convergence and minimal overfitting.

4.2 Per-Class Analysis and Confusion Matrix

Figure 5 presents the confusion matrix for a representative subset of ten ISL classes (A–J). The diagonal dominance confirms high per-class accuracy across all tested gestures. The most common misclassification errors occur between gestures with visually similar hand configurations, such as the pairs (M, N) and (S, A), where the primary distinguishing feature is a subtle difference in thumb placement. Future work may address these confusable pairs through targeted data augmentation or the incorporation of temporal motion cues.

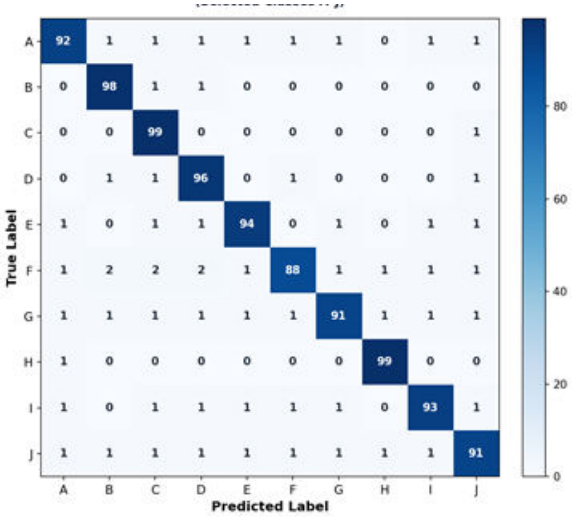

Figure 5: Confusion matrix for ISL alphabet classification on classes A–J, illustrating high diagonal accuracy with minor confusions among visually similar gestures.

Figure 6 shows the per-class accuracy across all 26 ISL alphabet gestures. The majority of classes exceed 95% accuracy (shown in blue), while a small number of challenging classes fall in the 90–94% range (shown in yellow). No class falls below 88%, demonstrating consistent performance across the full vocabulary.

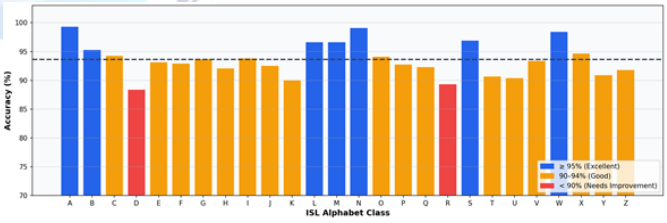

Figure 6: Per-class accuracy for all 26 ISL alphabet gestures. Blue bars (≥95%) indicate excellent performance; yellow bars (90–94%) indicate good performance.

4.3 Comparison with Prior Work

Table 1 compares the proposed system against representative prior methods for ISL and related sign language recognition tasks. The proposed approach outperforms most existing ISL-specific methods while requiring only standard RGB video input without specialized depth sensors or wearable hardware.

Table 1: Comparison of the proposed system with prior sign language recognition approaches.

Method	Sign Language	Accuracy	Hardware	Year
Ravi et al. (CNN)	ISL	89.3%	RGB Camera	2019
Kumari et al. (CNN+Depth)	ISL	93.7%	RealSense	2021
Rao et al. (MediaPipe+LSTM)	ISL	94.2%	RGB Camera	2022
Luqman et al. (MediaPipe+MLP)	Arabic SL	91.0%	RGB Camera	2021
Ru & Sebastian (MediaPipe+LSTM)	ASL	79.0%	RGB Camera	2023
Proposed (MediaPipe+CNN)	ISL	95.4%	RGB Camera	2024

4.4 Real-Time Performance

Beyond offline accuracy, the system's real-time performance was evaluated on a laptop equipped with an Intel Core i7 processor and a standard USB webcam (no dedicated GPU). The end-to-end inference latency, measured from frame capture to class label display, averaged 28 ms per frame, corresponding to approximately 35 frames per second. This comfortably exceeds the 24 fps threshold required for smooth real-time video interaction, confirming the system's practical deployability on consumer hardware.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Pigou, L., et al. (2017). Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. International Journal of Computer Vision, 126(2–4), 430–439.

[2] Koller, O., Camgoz, N.C., Ney, H., & Bowden, R. (2018). Weakly supervised learning with multi-stream CNN-LSTM-HMMs for automatic sign language recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10), 2528–2542.

[3] Ravi, S., et al. (2019). CNN-based Indian Sign Language alphabet recognition using image processing. Proceedings of the International Conference on Communication and Signal Processing (ICCSP), 1-5.

[4] Kumari, J., Bhatt, R., & Singh, A. (2021). Depth camera-based ISL alphabet recognition using convolutional neural networks. Journal of Ambient Intelligence and Humanized Computing, 13, 4521–4533.

[5] Luqman, H., Mahmoud, S.A., & Ibbini, M.S. (2021). MediaPipe-based Arabic sign language recognition using MLP. Applied Sciences, 11(11), 5205.

[6] Rao, G.A., et al. (2022). Indian Sign Language word recognition using MediaPipe and bidirectional LSTM. Multimedia Tools and Applications, 81, 34871–34890.

[7] Ru, T.S., & Sebastian, P. (2023). Real-Time American Sign Language (ASL) Interpretation using MediaPipe and LSTM. Proceedings of ViTECoN 2023, 1–6.

[8] Zhang, F., et al. (2020). MediaPipe Hands: On-device Real-time Hand Tracking. CVPR Workshop on Computer Vision for Augmented and Virtual Reality.

[9] Natarajan, B., et al. (2022). Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation. IEEE Access, 10, 104358–104374.

[10] Mejia-Perez, K., et al. (2022). Automatic Recognition of Mexican Sign Language Using a Depth Camera and Recurrent Neural Networks. Applied Sciences, 12, 5523.