Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 4 (2026)

A Systematic Review of Multimodal Representation Learning for Indian Sign Language Recognition Using Spatial Temporal and Transformer Based Approaches

Shivam Gautam[1], Aditi Sharma[2], Shubham Gautam[3]

[1]Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India

[2]Institute of Engineering and Technology, Lucknow, India

[3]B.N. College of Engineering and Technology, Lucknow, India

Abstract

In India, Indian Sign Language is the important means of communication for people from the Deaf and hard-of-hearing communities, with automated recognition providing the opportunities for overcoming communication problems and supporting simple human-computer interaction. The studies in the field have moved from the hand-crafted features and traditional machine learning to the use of CNN-based representation learning and temporal modelling, landmark and pose-based representations, multimodal learning and the use of Transformer-based technologies. Herein, we carry out a comparative analysis of 36 research articles dedicated to ISL and relevant sign language recognition techniques being focused on research datasets used, tasks, modalities, representations, architectures applied, strategies of fusion, results obtained and the limitations of the investigation. The findings of our analysis suggest a shift from traditional isolated signing classification techniques towards dynamic and multi-modal approaches capable of taking into account the timing of signs and the variability in the signing. While CNNs and transfer learning techniques seem to be working well for related tasks, new methods involving the use of language models can handle more complex patterns in signing. Nonetheless, problems with continuous recognition, generalization across signers, dataset diversity and benchmarking, robust multimodal fusion and computational efficiency are still a huge concern. Thus, according to our results, the integration of RGB and depth video, hand landmarks and body position along with effective temporal modelling seems to be the way to make ISL recognition systems more robust, generalizable and deployable in practice.

Keywords: *Indian Sign Language Recognition, Deep Learning, Multimodal Fusion, Hand and Body Landmarks, Pose Estimation, Transformer Networks*

1. Introduction

1.1 Background and Significance of Indian Sign Language

ISL plays a significant role in the lives of Deaf and hard-of-hearing individuals in India. Sign language is not just random gesturing - it is a complete language that consists of hand movements, hand shapes, body posture, facial expressions, as well as many other changes that do not involve hands [1],[2]. Because it is so complicated, researchers have come up with recognition methods that take into account not only hand gestures but also

parameters such as static images, movements in time and landmarks based on coordinates [3],[4],[5]. The increasing interest in automated ISL recognition systems is due to the communication barrier between ISL users and individuals who have not mastered sign language. The use of computer vision and machine learning in this situation allows for the translation of visual signs into formats that are easier to understand, like text or speech. The newest ISL systems rely on techniques like webcam-based recognition, facial marking extraction, deep learning classification and text-to-speech transformation to facilitate communication [5],[6],[7],[8]. Moreover, real-time detection and captioning can also serve as a means of detecting several signers in case of using computer vision systems [9].

Nevertheless, it still needs more complex techniques to recognize sign language than it takes to differentiate between isolated images of hands. The way a sign is understood depends on the configuration and motion of the hands, how they relate to the body in space, the sequence of movements timewise and non-manual components, such as movements of the face or lips [10],[11]. A number of recent studies on the topic of recognition of signs based on the Transformer-based recognition have emphasized the fact that signing process involves use of multimodal means and has shown significance of hand and finger signals and lip, face and body gestures of the signers [3],[4],[12]. One of the articles presented by Ghadami et al. proves the fact that hand, lips, wrist, arm and geometric details could be used in multi-stream Transformer-based framework to recognize Iranian sign language [3]. ISL uses a wide range of hand gestures and configurations to communicate language information. The alphabet of ISL is shown in figure1, which demonstrates the various hand shapes required in sign language.

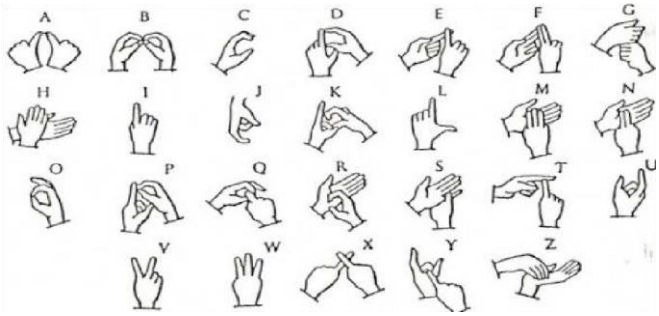


Fig. 1. Alphabet of ISL demonstrating the various hand configurations used to represent each letter [14].

1.2 Development of Automatic Sign Language Recognition

Based on the literature review conducted in this research regarding ISL and other similar sign language recognition systems, it is noted that the methods have evolved over time. The older studies predominantly focused on the feature engineering and traditional machine learning techniques such as SVM, KNN, PCA, SURF and classifier fusion [13],[16]. While these approaches reinforced the possibility of automatic sign recognition, the outcomes were still highly dependent on the design of features and the specificity of the sensors employed [14],[15]. The advent of CNN changed the direction of research to automatically learn visual representations. CNN along with transfer-learning architectures like VGG, ResNet and MobileNet achieved remarkable performance in many areas including static image and number recognition [6],[8],[17]. Katoch et al. achieved accuracy of 99.64% on a dataset of 36 classes having 36,000 images, while Sharma and Anand achieved recognition rates of 99% for digits and 97.6% for characters using their CNN [13]. The findings indicate the successful use of deep visual models with static recognition. Still, methods based on images cannot provide useful information about the processing of dynamic signs. To overcome this limitation, researchers began the process of the emergence of video-based approaches implemented by various algorithms, such as 3D-CNN and LSTM [19],[20],[21].

The INCLUDE [39] dataset that contains a total of 4,292 videos of 263 ISL signs became a very handy resource across the said video-based approaches focusing on the dynamic recognition in motion study and alternative varieties of representation. Another remarkable advancement is the application of structured hand and body landmarks acquired through MediaPipe, PoseNet and HRNet. Recent research has incorporated MediaPipe landmarks with light-weight neural networks for static recognition, transfer learning models and temporal models for dynamic recognition [6]. Studies on Pose-based methods have similarly utilized keypoint acquisition

with LSTM-based sequence processing to achieve real-time gesture-to-speech conversion [18]. Recently, multimodal and Transformer-based approaches emerged by fusing complementary RGB, hand, skeleton, pose and lip movements [3],[12]. Various architectures like TransMoVIF utilize dual-stream Transformers with cross-attention to combine visual and motion information further highlighting the growing trend of using complex multimodal systems in sign language recognition [4]. The below figure 2 illustrates this overall progression, tracing the shift from handcrafted and learned spatial representations to temporal, landmark/pose-based, multimodal and Transformer-based modelling. The development of ISL and SLR systems can broadly be viewed as a progression given below.

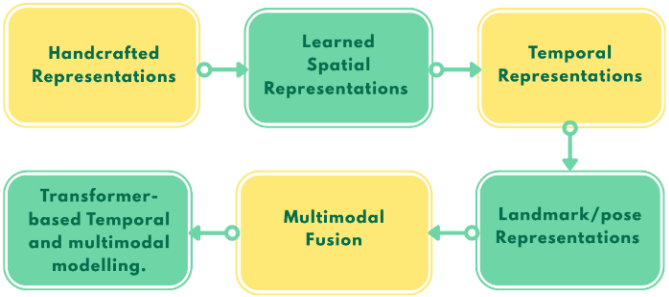


Fig. 2. Progression of ISL and SLR methods from spatial representations to temporal, landmark/pose-based, multimodal and Transformer-based techniques.

1.3 Challenges in ISL Recognition

1.3.1 Temporal and Signer Variability Challenge

The majority of current studies deal with fixed alphabets, numbers or single symbols, as recognition can be simply thought of as identifying certain visual shapes [6],[8]. In contrast to signing analysis, continuous signing is considerably more complex, as it involves temporal relations, transitions between signs, different motion patterns and other factors related to context. Studies employing LSTM, 3D-CNN, wearable IMUs and other time-oriented architectures show how it is important to explicitly model these temporal features [15],[18]. Different forming methods, styles as well as body postures and signing speeds create differences in performance of recognition systems. In addition, different aspects of the environment including background information, lighting level, clothing and camera parameters increase variability of vision-based systems [10],[11]. It becomes especially significant if we should deal with systems that need to generalize beyond individuals and conditions presented in the training models. Novel pose-based and Transformer-based approaches suggest native solutions to the given issue by employing structured spatial representations. For instance, Ghadami et al. used hand, lips, wrist/elbow and spatial data in the multi-stream transformer platform in order to develop the signer-independent model for Iranian Sign Language recognition [3]. Alkhoraif et al. researched appearance and pose information with the help of multimodal transformer fusion [12]. Various studies, while not specific to ISL, indicate suitable methodology in view of differences in the signer and in the representation.

1.3.2 Representation and Dataset Limitations

Every input mechanism presents its own challenges in sign language recognition. At pixel level, complex RGB video streams can capture intermediate images but also return unnecessary information such as background noise or anything that is not useful in the recognition process. At the same time, representations based on landmarks help categorize the positions of hands and body parts but might leave some data unrelated to their appearance out of the analysis. Thus, recent research shifts its focus to combinations of visual and landmark or positional representations [4],[5],[6],[12]. However, another significant limitation is found in the nature of datasets used in studies, with our comparison revealing vast differences in the size and the amount of data in terms of vocabulary, number of subjects, conditions of recording, modality of input and way of assessment. For instance, researchers analysed in the literature range from small custom datasets to large ones like INCLUDE [39] (4,292 videos for 263 signs), while some other datasets related to sign languages differ considerably in the number of elements. The found differences prevent drawing direct comparisons between accuracy values.

Consequently, performance results need to be interpreted within the specific dataset, task and evaluation setting in which they were obtained rather than treated as universally comparable measures [10],[11].

1.3.3 Computational and Deployment Challenges

With the advent of multimodal and Transformer model technology with many levels of complexity, richer representations are available, however they call for more computational power. Thus, real-time and resource-limited deployment becomes more challenging. In scientific works there has been a growing interest in addressing the accuracy-efficiency dilemma through lightweight CNN architectures, knowledge distillation, efficient detection models and landmark-based systems [5],[9]. In an example, researchers Luciana and others studied the process of obtaining lightweight Indonesian Sign Language models by applying the technique of knowledge distillation, while Ingoley and Bakal looked at various architectures like MobileNetV2 and bigger models in the realm of transfer learning when researching Indonesian Sign Language [8],[22]. In a similar vein, Raja and colleagues examined a number of YOLO variations for real-time ISL captioning and achieved low latency, thus evidencing that both the accuracy of models and their availability in terms of computational and latency requirements are essential for practical ISL recognition [9].

1.4 Motivation for the Review

Despite the progress made in the technology behind ISL recognition, we found that all previously reviewed studies of the topic differ significantly in terms of task specifications, input modalities, datasets and models used [10],[11]. The inconsistencies found in characteristics of the datasets and evaluation settings make it hard to compare accuracy measures provided in reviewed studies, especially when the studies under consideration deal with different sign languages, user populations, input modalities and types of tasks. The lack of standardization in methodology invites us to look at the field from a more comprehensive perspective. Rather than focusing only on isolated performance metrics, our synthesis examines the underlying progression from handcrafted features and conventional classifiers toward CNNs, temporal models, landmark-based approaches, multimodal fusion and Transformer architectures [3],[4]. By examining these developments collectively, we identify persistent limitations in continuous recognition, signer-independent generalization, dataset diversity, multimodal integration and computational efficiency. These observations provide the basis for identifying research gaps and emerging design principles for more robust and practical ISL recognition systems.

1.5 Aim and Objective of the Review

The main aim of this review paper is to present a well-organized along with a critical overview of the notable developments made in the field of Indian Sign Language recognition through deep learning methods. For this purpose, 36 research publications on Indian Sign Language were reviewed and research work done on other sign languages has been included wherever necessary since it helped in establishing comparison with the existing methodologies. The literature has been classified according to the underlying datasets, tasks related to recognition process, types of input used, feature representation strategies employed, models used, results obtained, drawbacks and information regarding the gaps in the research process. The paper maps the technological development of ISL recognition from traditional approaches to modern deep-learning systems. It analyses the present-day data landscape taking into account three basic components: datasets, vocabulary and signer population, covering input types such as RGB images and videos [15],[23], coordinates-based hand and body landmarks [5],[18] and wearable sensors [24]. Additionally, the paper looks into CNN applications [19],[25]as well as transfer-learning models [6],[8], architecture-based and lightweight models [9], multimodal fusion technologies [4],[12] and the use of Transformer-based models in static, isolated and continuous recognition tasks [3],[12]. Crucially, we do not consider reported accuracy alone to be enough to rank models. Alternatively, performance of the system is interpreted in the context of the dataset, the recognition task, number of signers, input modality and the evaluation situation described in each study. Based on this comparative analysis, we identify the recurring issues related to continuous recognition, robustness of the system, signer-independent generalization and standardization of datasets, multimodal integration and computational effectiveness. The outcomes of the analysis are employed later to formulate research gaps and future directions in practical ISL recognition.

2. Literature Review

2.1 Progression of Indian Sign Language Recognition

The automated system for identifying ISL has evolved through multiple processes. By examining the 36 pieces of research reviewed in this paper, it can be noted that in the course of time development has moved from using traditional techniques of machine learning and handcrafted features to other advanced methods including temporal modelling, landmark and pose representation, deep learning and modalities fusion as well as Transformer-based architecture [15],[16]. Progress in the field means the ability not only to understand signs in terms of their spatial quality but also in terms of temporal and spatial characteristics. [18],[19].

Table 1. Comparative analysis of recent studies on ISL and SLR.

Author (Year)	Technology Used	Performance	Dataset	Key Findings
Ghadami et al. (2027) [3]	Multi-stream Transformer Encoder using early + late fusion, with MediaPipe for hand/lip keypoints, YOLOv5m for hand/face detection, word embeddings + multi-task learning and GA for ensemble optimization.	Top-1 Accuracy: 90.2%; Top-5: 93.1%. Late fusion: 88.8%; Early fusion: 85.4%.	ISLR 101, Total 101 words, 4,040 videos, 11 signers; 600×800 resolution, 25 FPS.	Combines hand, lip, wrist/elbow, hand-distance and angle features for signer-independent recognition. Uses early and late Transformer fusion and a learnable ensemble to improve recognition. Adding lip features significantly improves performance when combined with hand features.
Renjith et al. (2026) [4]	Dual-stream Transformer using motion trajectory + visual features.	Include-50 Acc : 96.80% SMILE-DSGS Acc : 93.95%	Include-50 Dataset SMILE-DSGS Dataset	Proposes (TransMoVIF). Combines motion + visual (RGB) features. Uses cross-attention fusion mechanism.
Patel et al. (2026) [5]	MediaPipe Hand Landmark Detection with a Sequential ANN model for static ISL gesture classification.	Accuracy : 98.00%	Custom ISL Dataset containing 26 alphabets + 7 words; 75% training, 25% testing.	Proposes a real-time vision-based ISL recognition system using MediaPipe. Uses 21 hand landmarks with normalized (x, y, z) coordinates are used as features and a lightweight Sequential ANN for classification.
Shilpa Ingoley et al. (2025) [6]	Transfer Learning + hand landmarks. CNN models (VGG16, VGG19, ResNet50, MobileNetV2), Google MediaPipe, Adam Optimizer, TensorFlow 2.15, Hand Landmark	Validation accuracy: VGG16 99.76%, MobileNetV2 99.44%; F1-score: VGG16 0.995, MobileNetV2 0.99.	ISL_Num14_AS3N24: 6,300 images, 14 classes (0-9, alternate 9, Thousand, Lakh, Crore), 5 signers, webcam capture, 224×224×3,	Recognizes ISL numerals and converts them to text and speech. Handles left, right and ambidextrous signers, selfie-mode capture, gloved hands and signs with multiple gestures. VGG16 is slightly more accurate, while MobileNetV2 is about 3x smaller (42.8 MB vs 130.1 MB) and trains about 2x faster (6.1 vs

	Detection, Image Preprocessing & Padding Techniques		80:20 split, no augmentation	11.8 min).
Pranjal Gogoi et al. (2025) [7]	Lightweight CNN + PyTTSX3 for text-to-speech conversion	Accuracy : 98.55% (ASL), 99.64% (ISL) Inference Time : 76.38 ms	ISL Dataset ASL Dataset	CNN-based gesture recognition model. Integrated with text-to-speech (pyttsx3). Works on Indian & American datasets.
Shilpa Ingoley et al. (2025) [8].	Transfer learning with four pretrained CNNs. MediaPipe palm detection and 21 hand landmarks for cropping; Adam optimizer, early stopping.	Val. Accuracy : 99.67% (VGG16), 99.69% (MobileNetV2), 98.72% (ResNet50), 99.69% (mobileNetV2)	Custom ISL Alphanumeric Dataset created using webcam and MediaPipe hand landmarks.	Recognizes static ISL alphabets and numerals using hand landmarks and transfer learning. MobileNetV2 performs best overall: highest validation accuracy, smallest size and few epochs (7). ResNet50 has the lowest accuracy.
Rohit Raja et al. (2025) [9]	YOLO (v5, v6, v7, v8, v9, v10, v11, YOLO NAS) for real-time ISL detection	mAP : up to98.2% YOLO NAS : mAP@0.50 = 0.982–0.983 YOLO 11 : mAP@0.50 = 0.978 YOLOv8 : mAP@0.50 = 0.975 YOLOv9 : mAP@0.50 = 0.974 YOLO version 5 : mAP@0.50 = 0.969 YOLO version 10 : mAP@0.50 = 0.960 Real-time latency : 2.36–7.58 ms Speed : 12–25 FP	Image-ISL-PICT Dataset (self-created)	Compared YOLO v5, v6, v7, v8, v9, v10, v11 and YOLO-NAS. Dataset: 115 ISL classes, 10 signers. Best model: YOLO-NAS (mAP~0.982). Supports real-time multi-person detection. Works with live webcam input.
Lamia Trabelsi et al. (2026) [10]	CNN, RNN, SignBERT, Transformer, CA-SignBERT, YOLOv5 / YOLOv6, LSTM. HMM, SVM, CNN-LSTM Hybrid Models, Graph Convolution Networks (GCN), MediaPipe,	CNN + TIM : 99.8%. Hybrid InceptionNet: 98.46% SVM-based recognition : 98.2%. YOLOv6 static sign recognition : 96% GCAR model: 94.10%. VGG16/MobileNet V2 based ISL	WLASL, MS-ASL, UTSL, RWTH-PHOENIX-Weather 2014 , CSL-Daily, LSA64, SIGNUM, KArSL.	Comprehensive survey of SLR techniques (ML + DL). Identifies key challenges: occlusion, lighting, dataset variation. Explains manual + non-manual features (hands + face + body). Highlights need for real-time SLR systems. Provides taxonomy of SLR approaches.

	Transfer Learning Approaches.	model: 73% Spatio-temporal CNN: 84.5%		
Reza Jalayer et al. (2026) [11]	Covers hand detection, segmentation, gesture recognition.	N/A	ASL, LSA64, ISL, Korean SL Isolated gesture dataset, WLASL Dataset RWTH-PHOENIX-014T, ISL-CSLTR.	Covers hand detection, segmentation, gesture recognition. Reviews CNN, RNN, Transformers, GNNs. Highlights importance of vision-based systems in HRI. Applications in robotics & assistive tech.
A. A. Alkhoraif et al. (2026) [12]	Swin Transformer (appearance + pose fusion)	Top - 1 Accuracy: WLASL: (93.51% -88.53%) MSASL: (92.56% -91.89%) ASL Citizen: (86.53% -84.23%)	WLASL100, LASL300, MSASL100, SASL200, ASL Citizen100, ASL Citizen300.	Two-stream multimodal architecture. Converts pose landmarks into RGB images. Achieves high accuracy (93.51% WLASL). Captures spatial + temporal features effectively.
Katoch et al. (2022) [13]	SURF, Bag of Visual Words (BoVW), Mini-Batch K-Means, SVM and CNN with skin-color segmentation and bg. subtraction.	CNN : 99.64% SVM : 99.17% Precision : 99.57% Recall : 99.57%.	Custom ISL Dataset; 36,000 images, 26 alphabets + 10 digits, 3 signers, 1000 sign.	Proposes a real-time ISL recognition system. CNN achieves the highest accuracy. Supports speech output and speech-to-sign conversion. Handles both single- and double-handed gestures.
Kakoty et al. (2018) [14]	Data glove-based ISL/ASL recognition using hand kinematics captured through wearable sensors.	Not Reported.	Data-glove based ISL & ASL alphabets and numerals dataset.	Uses wearable glove instead of cameras. Captures finger joint kinematics. Supports ISL and ASL alphabets and numerals. Eliminates lighting dependency.
Tripathi et al. (2015) [15]	Gradient-based Key Frame Extraction, Orientation Histogram (OH), PCA and distance-based classifiers.	Best Accuracy: 93% (Euclidean & Correlation Distance).	Custom Dataset; 10 sentences, 5 participants, 10 repetitions per sentence.	Continuous ISL sentence recognition 640×480 resolution. Uses key-frame extraction. OH (36 bins) + PCA improves recognition. Euclidean and Correlation classifiers achieve the best performance.
Krishnaveni et al. (2012) [16]	Feature extraction with KNN, MLP, SVM and Bayes Aggregation	Best Accuracy: 92.8571% (KNN + SVM with Bayes aggregation).	Self-created View-based ISL Dataset; 60% training, 40%	Proposes classifier fusion for ISL recognition. Bayes aggregation outperforms majority voting. KNN + SVM yields the highest

	classifier fusion.		testing.	accuracy. Effective for isolated sign recognition.
Sharma & Anand (2021) [17]	Comparative study of pre-trained CNNs with Transfer Learning and a Proposed 3-layer CNN.	Proposed CNN : 99.0% (Numerals), 97.6% (Alphabets) Best Pre-trained (ResNet152V2) : 96.2%, 90.8%.	Public ISL Dataset containing static alphabets and numerals.	Evaluates multiple transfer learning models. Proposes lightweight CNN. Studies optimizer and hyper parameter effects. Proposed CNN outperforms pre-trained models.
Shubham Shetty et al. (2024) [18]	Pose Estimation (PoseNet) + LSTM (RNN) for real-time gesture-to-text.	Accuracy: 98%	INCLUDE Dataset (Indian Sign Language Dataset) INCLUDE	Uses PoseNet for keypoint extraction. LSTM for sequence prediction. Real-time gesture-to-text conversion.
Dushyant Kumar Singh (2021) [19]	3D-CNN for dynamic ISL gesture recognition using spatio-temporal video features.	Accuracy : 88.24% Precision : 88.69% Recall : 88.88% F1-score : 88.52%.	Self-created Dynamic ISL Dataset; 20 gestures, 10 volunteers, 2400 videos, 162,840 frames.	Learns spatio-temporal features directly from videos. Introduces a dynamic ISL dataset. Eliminates handcrafted feature extraction. Suitable for dynamic gesture recognition.
Gupta et al. (2020) [20]	Continuous ISL classification using Wearable IMUs, t-LeNet, Modified t-LeNet and MC-DCNN.	MC-DCNN: 83.94% Modified t-LeNet: 81.62% t-LeNet: 79.70%.	Self-created IMU Dataset; 11 continuous sentences, 15-word vocabulary, 10 volunteers, 1100 samples.	Compares three CNN architectures. Uses wearable IMUs instead of cameras. Modified t-LeNet reduces parameters. MC-DCNN achieves the highest accuracy. Demonstrates efficient wearable-based continuous ISL recognition.
Suri et al. (2019) [21]	Continuous ISLR using Wearable IMU, 1D Deep Capsule Network (CapsNet) and Game Theory (Nash Equilibrium).	CapsNet : 94.0% CNN : 87.99%.	Self-created Continuous ISL IMU Dataset.	Introduces a 1D Deep Capsule Network. Uses custom wearable IMUs. CapsNet outperforms CNN. Effective for continuous ISL recognition.
Devan Luciana et al. (2025) [22]	CNN + Transformer models (EfficientNetB5, Swin Transformer)	Best teacher model : EfficientNetB5 (Accuracy : 96.94%) Best lightweight model : ShuffleNetV2 (89.87%)	BISINDO (BahasaIsyarat Indonesia) Alphabet Dataset 14,406 images, 26 classes (A–Z), BISINDO Dataset Repository	Applies Knowledge Distillation (KD) to create lightweight models. Hybrid KD (DKDF) improves performance significantly. Achieves good balance between accuracy & efficiency.

Rao et al. (2018) [23]	Selfie Video-based continuous ISLR using Sobel Edge Detection, Morphology, MDC, ANN and Adaptive Thresholding.	ANN: 90.00% WMS MDC: 85.58% WMS.	Smartphone selfie video-based continuous ISL dataset.	Proposes smartphone-based ISL recognition. Improves Sobel segmentation. ANN outperforms Minimum Distance Classifier. Suitable for mobile devices. Enables real-time continuous recognition.
Gupta et al. (2021) [24]	Wearable sEMG + IMU sensors with Multi-label Classification using lexical attributes and symmetry features.	Classification Error: 2.73% (Proposed) 6.22% (Hierarchical Classification).	Self-created ISL Dataset; 100 isolated signs, 10 signers, multiple sEMG and IMU sensors.	Introduces multi-label classification. Uses lexical and symmetry features. Reduces error propagation. Improves over hierarchical classification. Robust wearable-based ISL recognition
Tak et al. (2026) [25]	CNN and LSTM for static/dynamic gesture recognition, with MediaPipe hand landmarks for feature extraction and motion-based routing.	CNN Acc.: 95.0 CNN F1: 97.0 LSTM Acc.: 96.0 LSTM F1: 96.0	Composite ASL + ISL dataset: 61 static classes (120,000 images) and 50 dynamic classes (5,000 sequences). Static: 70:15:15 Dynamic: 80:20.	Uses MediaPipe hand landmarks to support real-time temporal feature extraction. Motion-score routing directs static gestures to CNN and dynamic gestures to LSTM. Real-time 15–25 FPS performance.
Mutlu et al. (2026) [26]	NLP-based Sign Language Processing.	POS (93.85%) English to ASL: BLEU-1: 93.26	RWTH-PHOENIX-Weather 2014 / 2014T, How2Sign, CSL-Daily WLASL, MS-ASL, AUTSL, ASLLVD.	Covers translation, POS tagging, NER, sentiment analysis. Highlights dataset scarcity and linguistics. Suggests need for interdisciplinary (CV/NLP/linguistics) approach.
Yerimbetova et al. (2026) [27]	MediaPipe + Temporal Convolutional Network (TCN) + Transformer	Accuracy: 75% Val. Acc.: 72%	KRSL Dataset	Uses MediaPipe for 225 keypoint extraction. Handles temporal + spatial gesture features. Suitable for low-resource SL datasets.
Venkatesh et al. (2026) [28]	Hybrid CNN (ViT-YOLOv5 + SIFT + EfficientNetB7 + 1D & 2D CNN) optimized using FAOA	D1 Acc: 93.88% D2 Acc: 92.63% D3 Acc: 94.57% D4 Acc: 93.88%	ISL Dataset, ISLRTC Referred Dataset, Gesture Recognition Dataset, SL Gesture Images	Uses hybrid model. Combines 1D & 2D CNN for feature processing. Achieves superior performance over baseline models.

			Dataset	
W. Ismaiel et al. (2026) [29]	CNN + ResNet50 + CED Method 1: ASLR_CNN + CED Method 2: ResNet50 + CED	M1 Accuracy: ArASL: 97.14% AASL: 89.49% M1 Accuracy: ArASL: 96.88% AASL: 86.12%	AASL(Kaggle) ArASL(Mendeley)	Canny Edge Detection + CNN. Improves feature extraction using edge-based pre-processing. Works on Arabic SL datasets. Reduces computational cost and improves robustness.
Umar Ashfaq et al. (2026) [30]	Vision Transformer (ViT) with attention mechanism	Accuracy: USL: 99.70% ArSL2018: 99.52% ASL 1st one: 99.44% ASL 2nd one : 99.84%	USL Dataset (40 Urdu sign alphabets), ArSL2018, ASL 1st version and 2nd version	Captures long-range dependencies (LRD). Handles visually similar gestures effectively. Works on multiple sign languages (ASL, Urdu, Arabic).
Md. F. I. Pranto et al. (2026) [31]	MediaPipe processed + raw data	N/A	ASL-HG (American Sign Language Hand Gesture Image Dataset)	Provides 36,000 images (36 classes) dataset. Includes alphabets (A to Z) + digits (0 to 9). Balanced dataset (10 subjects, diverse conditions). Useful for training robust DL models.
Sharath S R et al. (2025) [32]	Seq2Seq + HamNoSys + SiGML + Keypoint Estimation (Multi-language translation)	HR Acc: 93.68% HPR Acc: 99.28% HOR Acc: 99.28% Seq2Seq BLEU-1 score: 0.857 ROUGH-L Score:0.88	Custom datasets created using Google MediaPipe key-point estimation for: HR, HPR, HOR using ISL, ASL, LSF.	Uses Seq2Seq model with HamNoSys representation. Converts between different sign languages. Uses SiGML for avatar generation.
Abhishek Tripathi et al. (2025) [33]	IoT-based glove + Flex Sensors + Arduino UNO (Hardware-based recognition)	Accuracy: 85% (simple gesture), 75% (complex gest.)	Custom glove-based gesture dataset collected using: Flex sensors, Arduino UNO-IoT.	Uses IoT glove with flex sensors + Arduino. Converts gestures into text in real-time. Low-cost and portable system.
Harapriya Kar et al. (2025) [34]	Nth Layer Hierarchical Bidirectional LSTM (NHBLSTM) + Co-attention mechanism	NHBLSTM Acc: 99%	ASL (American Sign Language) Alphabet Dataset 29 classes, JPEG dataset.	Proposes NHBLSTM with co-attention. Reduces error accumulation in sequence learning. Improves temporal modelling.
Lincolin	Google MediaPipe	Achieved 99% test	Khasi SL	Applies CLAHE preprocessing

Amory Ranee et al. (2025) [35]	Hand Tracking + CLAHE preprocessing + Feedforward Neural Network (FNN) + TensorFlow/Keras/ OpenCV	accuracy with 0.0355 test loss on 3451 test images.	Dataset (Self-Created Dataset). 18 static two-hand Khasi alphabet signs: A, B, D, E, G, I, Ï, K, M, N, Ñ, NG, P, R, S, T, U, W	for better image visibility. Developed Feedforward Neural Network (FNN) for Khasi sign recognition. Created custom dataset with 24,000 static sign samples.
Mishra et al. (2025) [36]	Random Forest, LSTM, Gemini-1.5-Flash LLM, MediaPipe + OpenCV	RF- 99.98% accuracy; LSTM: 93% accuracy; F1-score up to 0.56 for selected words.	Kaggle ISL: 42,000 images, 35 classes; Custom dataset: 18 words each have 30 video instances.	Integrates static and dynamic ISL recognition. LSTM + LLM generates contextually meaningful sentences. Provides real-time translation along with learning, flashcard and quiz modules.
Renjith S et al. (2024) [37]	Spatio-temporal features + Time Distributed CNN (TD-CNN) + LSTM	Accuracy CSL: 90.87%, ArSL: 89.46%	Arabic SL Dataset (ArSL) Chinese SL Dataset (CSL)	Uses spatio-temporal features from video sequences. Captures both spatial and motion information. Works on CSL and ArSL datasets.
Chaudhari et al. (2024) [38]	SNN + ResNet-based architecture + Transfer Learning	Improvement over state-of-the-art, 25% lower activations	WLASL300 (Word-Level American Sign Language Dataset)	Energy-efficient Spiking Neural Networks (SNN). Handles temporal data effectively. Improves accuracy by 2.7% over SOTA.

2.1.1 The Classical and CNN-Based Recognition

Early approaches to ISL Recognition framed the issue of detection in the light of the problem of detection and of classification. Hand-crafted methods such as PCA, SURF, BoVW, orientation histograms and other kinds of image or motion descriptors were combined with classifiers like SVM, KNN, MLP and Bayesian aggregation [15],[16]. As distinct from that, the sensor-based platforms apply different technology (data gloves, IMUs, sEMG sensors) for capturing the motion and the kinematic data [14],[21],[24]. The findings showed that it was possible to apply the technology for ISL recognition and simultaneously demonstrated that the ability to recognize depended on manually created features, controlled conditions for the data obtaining and specialized equipment. With the advent of CNNs, the focus has turned towards importing the power of learned visual representations. CNN and transfer-learning applications including VGG, ResNet, MobileNet and various other deep models have been used in ISL alphabets, numbers and hand gestures [6],[8].

Katoch et al. achieved 99.64% accuracy using an ISL dataset that included 36 classes and 36,000 images while Sharma and Anand announced 99.0% accuracy in recognizing numbers and 97.6% accuracy in recognizing alphabets with their CNN architecture [13],[17]. Ingoley and Bakal also reported above 99% accuracy with their transfer-learning model on their unique ISL alphanumeric dataset [6]. All these results highlight the suitability of CNN representation in static recognition. The strong performance of the models in the area of static classification is accompanied by a major drawback. A single image provides the necessary information for determining the locality of a sign; however, it does not show how a sign is produced from the movement of hands, nor does it represent the transition from one sign to another or any temporal progression of the signing process. The significance of this limitation grows when the focus of research shifts from single gesture classification to the area of dynamic and continuous recognition.

2.1.2 Temporal and Landmark-Based Recognition

The use of videos instead of static images has led to the understanding of the significance of features as regards time, thus encouraging scholars to move to the use of not only 3D-CNNs, LSTMs, Bi-LSTMs, temporal CNNs, capsule networks, but also attention-based sequence models [19],[20],[21]. For instance, Singh et al. have used a 3D-CNN in order to learn spatio-temporal patterns out of ISL dynamic videos and achieved an accuracy of 88.24%. In addition, previous studies on recognition, such as continuous recognition experiments, have preceded this research in the area of key-frame extraction and temporal modelling, while theories of motion by means of IMUs show how to capture movements without cameras [23],[24]. The INCLUDE [39] dataset has been a crucial source for studying dynamic ISL recognition and other alternatives since it includes 4,292 videos and 263 signs [18],[39]. In a recent publication, Shetty et al. applied pose recognition and LSTM sequence modelling as a method for real-time conversion of gestures into text, using the INCLUDE dataset to demonstrate combining the two methods [18].

A further significant innovation is the use of structured hand and body landmarks via technologies like MediaPipe, PoseNet and HRNet. In their article, Patel et al. talked about 21 three-dimensional hand landmarks as data given to a sequential ANN, producing 98% accuracy in their static ISL dataset [5]. Ingoley and Bakal also used MediaPipe hand landmarks in their ISL recognition pipeline [6],[8]. Landmark-based representations can minimize the amount of raw visual information processed by a model and deliver clear structural information on the configuration of a hand. However, such representations in the same may miss the visual and contextual information. This property leads to the idea of integrating landmark representations with visual information rather than treating them as rivals.

2.1.3 Multimodal and Transformer-Based Recognition

Current research is increasingly more focused on the incorporation of complementary modalities which include various features like RGB display, skeletal movements, hand reference points and body position [3],[4],[12]. This marks a significant change from a previous approach wherein a single form was asked to capture the features of sign language. Instead of looking for a single representation, the multimodal systems look for ways to include complementary spatial, temporal and structural information. TransMoVIF is an example of this trend and utilizes both visual RGB signals and movement trajectory through a dual-stream Transformer and cross-attention fusion technique [4]. The datasets for sign languages used in this survey showed that the model achieved 96.80% accuracy on INCLUDE-50 and 93.95% on SMILE-DSGS [4]. Other modern research documented the application of fusion of appearance and pose features by Swin Transformers [12]. Even though research deals with sign languages other than ISL, it may be useful from a methodological viewpoint when it comes to understanding the way in which multimodal representation may solve the problem of shortcomings of monomodal recognition.

Transformer-inspired models are starting to be present at various stages of recognition pipelines. Vision Transformers have been considered for establishing long-range spatial dependencies for static sign recognition [30], while multi-stream Transformer models have been used to facilitate integration of information related to hands, body and lips in more advanced recognition cases [3]. Ghadami et al. demonstrated the effectiveness of a multimodal framework with additional hand and lip features, although their study focused on Iranian rather than Indian Sign Language [3]. Overall, these observations reflect a gradual shift from single visual classification systems toward approaches that jointly capture spatial appearance, temporal dynamics and the heterogeneous nature of signing. Figure 3 summarizes this progression:

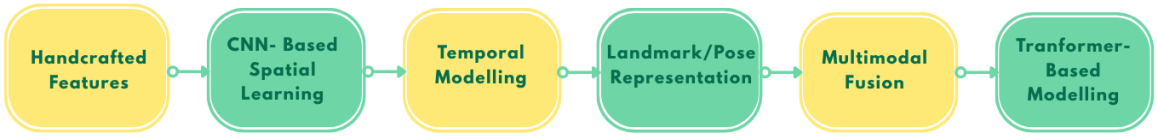


Fig. 3. Evolution of ISL/SLR methods from handcrafted and CNN-based features to temporal, landmark-based, multimodal and Transformer-based approaches.

2.1.4 From Recognition to Translation and Deployment

The Mover has resulted in a growing curiosity to research not just isolated class predictions, but also outputs related to text, speech and language. Ongoing contributions in this direction include ongoing research that focuses on ISL-to-text/speech conversion, gesture-to-speech interaction in real time and continuous sign-to-text conversion [33],[36]. Hence, ISL research has been moving forward from the research on gestures classification to sign recognition and translation as well as its application in practice with much more attention being paid to multimodal, time-aware and computationally efficient architectures [3],[4],[9],[25].

2.1.5 Summary of the evolution

The problem of research has evolved from presenting hand and visually crafted features of sign language computing to studying the spatiotemporal and structured representations of signs [15],[16]. Convolutional neural networks improved the learning of spatial features of signs and sequence models have become essential for understanding sequences, whereas models based on landmark detection, positioning, or multimodal transformer began being used for integration of additional information [4],[5]. The main issue faced is not just increasing the classification accuracy of the systems but forming the systems that are able to integrate the features of all above-mentioned types while solving such issues as temporal variations, great diversities of signers, environmental conditions and practical obstacles [3],[9]. According to the literature analyzed, the development of sign language recognition can be summarized as follows.

Table 2. Evolution of ISL/SLR recognition approaches from handcrafted features to deep learning-based spatial and temporal modelling.

Stage	Primary representation	Typical models	Main objective
Early approaches	Handcrafted/sensor features	SVM, KNN, MLP	Feature-based classification
CNN era	RGB images	CNN, VGG, ResNet, MobileNet, EfficientNet	Spatial representation learning
Temporal era	Video sequences	3D-CNN, LSTM, Bi-LSTM, temporal CNN	Motion and sequence modelling
Landmark/Pose era	Hand/body keypoints	FNN, LSTM, TCN	Structured motion representation
Multimodal era	RGB + landmarks/pose/sensors	Multi-stream models	Complementary information
Transformer era	Visual + structural sequences	ViT, Swin, Transformer encoder/decoder	Long-range dependency and fusion
Emerging deployment direction	Multimodal real-time streams	Lightweight CNN, YOLO, SNN, efficient Transformers	Robust and practical recognition

2.2 Datasets Used in ISL/SLR Research

Sign language recognition relies on the datasets on which it is based since aspects such as their size, vocabulary, signer diversity, acquisition conditions, modality and task definition are critical for the building and assessment of a sign language recognition model [10],[11]. In the course of analysing the current research, i.e. 36 reviews, static image datasets, dynamic video datasets, custom datasets, landmark-based relations and information based on wearing sensors from Indian and various sign languages were discovered. Although these datasets indicate the large quantity of research carried out in the field, comparison of performances is limited.

2.2.1 Dataset Diversity and Characteristics

Datasets comprised of static images are still frequently utilized in the domains of character, digit and isolated-sign recognition (for instance, Katoch et al. created a custom ISL dataset that consists of 36,000 photographs coming from 36 distinct classes featuring three individual signers) whereas Sharma and Anand focused on static alphabet recognition. While such datasets are valuable for developing spatial recognition algorithms, they hold little relevance when it comes to conveying temporal information. In recent years, video datasets have become increasingly crucial for dynamic and continuous recognition. The INCLUDE dataset is very important in ISL research, consisting of 4,292 clips and covering 263 signs [18],[39]. Its use has allowed inspecting temporal models and representations such as the MediaPipe and the use of features that are based on poses [18]. Other datasets such as ISLR101, which contains 4,040 samples of 101 words, can provide observations for methodology comparisons in large vocabulary recognition [3]. The research works show that a bigger set of samples does not always result in larger vocabulary or more complex tasks. Researchers still utilize custom datasets due to their flexibility in designing datasets around their experimental needs. Unfortunately, the differences in signer populations, vocabulary, recording settings, backgrounds, annotation techniques and train-test procedures complicate any comparison of different studies [10],[11]. Landmark-based representations offer another dimension to datasets as 21 three-dimensional hand landmarks or other MediaPipe-based representations can be used [5],[6],[8]. While such representations allow reduction of input dimensionality and provide certain types of structured data, the effectiveness of recognition can depend on landmark detection quality.

2.2.2 Major Dataset Categories and Benchmarks

The main datasets contained in the literature are detailed in Table 3. The analysis comprises the ISL datasets such as INCLUDE and a number of other custom datasets and some related sign language datasets including WLASL, MS-ASL, ASL Citizen, ArASL, AASL, ASL-HG and ISLR10[41]-[44]. The mentioned datasets are not assumed to be the evidentiary basis for ISL; therefore, they can't be treated as evidence, nevertheless they do provide a methodological basis for representation, architecture and evaluation that are applicable to ISL. The datasets collected by wearables using data gloves, IMUs and sEMG supply concrete results of hand movements and muscles work [14],[21],[24], but their reliance on complex gadgets leads to the limitations in both spontaneity and availability compared to camera-based approaches leading to the rise of interest in the vision-based landmarks and poses representations [5],[6],[18].

Table 3. Comparison of major ISL/SLR datasets in terms of data type, size, vocabulary and limitations.

Dataset Name	Language	Dataset Type	Size	Classes / Vocabulary	Key Limitation
ASL-HG [41]	American Sign Language (ASL)	Image	36,000 images	36 classes	No dynamic gestures
ISL_Num14_AS3N24 [6]	ISL	Image	Custom dataset	Alphabets + numbers	Limited vocabulary
WLASL [12],[38]	ASL	Video	2000+ glosses	Large vocabulary	Low top-10 accuracy
Custom ISL Dataset [8]	ISL	Image	21,000 images	35 classes	No sentence-level gestures
CSL [37]	Chinese Sign Language	Video	360 videos	Multiple gestures	Small dataset size
ArSL [37]	Arabic Sign	Image	7857 images	Alphabet	Moderate accuracy

	Language	/Video		gestures	
KRSL-2025 [27]	Kazakh Russian Sign Language	Video	43,250 samples (~44 hrs)	173 sentences	Low-resource dataset
Gesture Recognition Dataset [28]	Mixed	Image /Video	Not specified	Gesture classes	Lack of standardization
Sign Language Gesture Image Dataset [28]	Mixed	Image	Not specified	Gesture classes	Limited benchmark usage
AASL [29]	Arabic Sign Language	Image	7857 images	31 Classes	Dataset imbalance
ArASL [29]	Arabic Sign Language	Image	54,049 images	32 Classes	Similar gesture confusion
Image-ISL-PICT [9]	ISL	Image	3580 images	115 classes	Small samples per class
INCLUDE [18]	ISL	Video	4292 videos	263 word-level signs	Limited real-world diversity
INCLUDE50 [4] [18]	ISL	Video	958 videos	50 signs	Smaller subset of INCLUDE

2.2.3 Findings and Dataset-Related Research Gaps

An in-depth analysis of the recent data trend uncovers many valuable insights. For instance, several studies indicate extremely high recognition rates for static and isolated sign-language tasks [13],[17]. Even if such results help prove the effective spatial recognition mechanisms, this would only mean poor performance in continuous signing tasks that include both temporal dependencies and transitions between signs [20],[21]. Video-based resources like INCLUDE prove especially beneficial for exploring temporal recognition; however different vocabulary size, signing community, filming conditions and evaluation methods still hinder standardization [18],[39]. In a wider context, our comparison reveals the fact that limited data diversity is one of the main issues throughout the literature. The usefulness of a given dataset relies not only on how many samples it has, but also how broad the vocabulary used in it is, how diverse the participants are, how varied environments the data are taken in, the recording process conditions and the robustness of the adopted evaluation protocol [10],[11]. Therefore, fundamentally, the gap related to data is not only the absence of samples, but also the lack of standardized and real-world ISL datasets necessary for a reliable assessment and development of truly independent approaches to sign language recognition.

2.3 Classical Machine Learning Approaches

Initially, the research on ISL recognition relied heavily on the process of extracting features by hand to later classify them through traditional means. The data which comes from visual or other Sensors is processed into crafted patterns which can be classified by well-known classification methods based on the algorithms like SVM, KNN, MLP and many others [13],[14],[15],[16]. These approaches led to very solid fundamentals before the application of deep learning in ISL field.

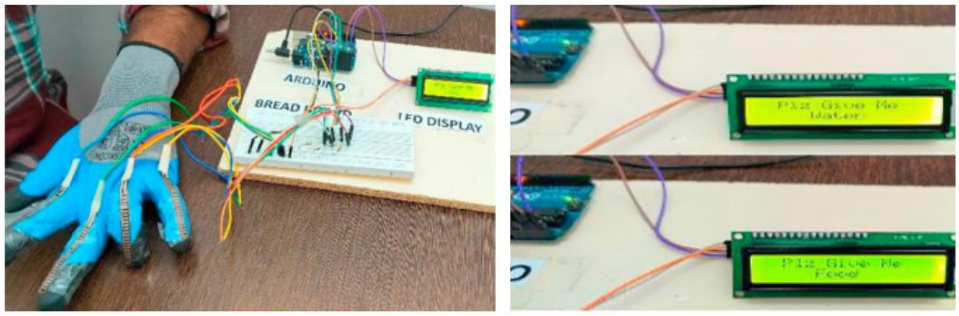
2.3.1 Handcrafted Features and Conventional Classifiers

Krishnaveni and Radha studied KNN, MLP and SVM as classifiers using image features making use of Bayesian aggregation and majority voting. Their most effective method that combined KNN and SVM with the help of Bayesian aggregation achieved 92.8571% accuracy thus showing the promise of classifier fusion in ISL

recognition [16]. Katoch et al. further attempted to move towards learned visual representations by combining SURF descriptors and BoVW with SVM and CNN classifiers. Their dataset consisted of 36,000 images from 36 classes and they achieved 99.17% accuracy with their SVM classifier and 99.64% with the CNN classifier [13]. The primary benefit of these methods is their comparative ease of use and success in constrained environments. Handcrafted features can explicitly recognize edges, shapes, patterns and positions of hands, while standard classifiers use simpler representations of features [13],[14],[15],[16]. Still, their efficiency depended on how good and sturdy the selected features were. Lighting, background, hand position and movement could impact their acting ability in different signing situations. Hence these shortcomings led to the introduction of deep learning methods in which representations could be derived more effortlessly from visual data.

2.3.2 Wearable and Sensor-Based Recognition

One of the simultaneous research paths employed wearable devices for taking hand and body movements' data. The research carried out in the collected literature made use of data gloves, IMUs and sEMG. Kakoty and Sharma utilized a data glove for obtaining hand kinematics, whereas Gupta and Kumar studied sEMG and IMU signals for ISL recognition [14],[24]. Other scientists studied a continuous recognition using the models based on IMUs and CNN and capsule network architectures [20],[21]. The above-given approaches enable the generation of structured data on the motion as well as lowering the requirement for visual conditions [14],[21],[24]. Still, in comparison to visual systems, the custom equipment imposes extra requirements with regard to the setup of devices and their usage by users. The practical limitations, together with the inflexibility of engineered features led to the development of deep learning techniques. However, technology-based studies showed the importance of the collection of the motion and kinematic data, which is still valid in modern localization and landmark recognition systems [4],[12],[18].



(a) (b)

Fig. 4. Prototype IoT glove for ISL recognition: (a) flex sensor and Arduino UNO circuit, (b) real-time gesture-to-text output on LCD display [33].

2.4 CNN and Transfer-Learning Approaches

The discovery of CNN technology has altered the way sign languages are recognized. CNNs have taken the place of the traditional way of using human-designed features. CNNs can learn features such as the shape, position, silhouette and image of the hands easier and better compared to other techniques for recognizing fixed ISL, [13],[17]. This change is explained by the authors of the research that they did. Katoch et al. tested both SURF manual features and bag-of-visual-words method against CNNs with the results of about 36,000 pictures of 36 types of ISL. The accuracy of using CNN is 99.64% while SVM works only 99.17% [13]. This is a good example of growth in terms of effectivity of CNN as the above-mentioned method still produces results. The popularity of transfer learning has increased the value of CNN-based recognition.

Several networks have been used for ISL recognition, as transfer learning enables the models to use previously learned visual representations for signs [6],[8],[17]. As an example of energy-efficient architectures, MobileNet is particularly suited for webcam-based and low-resource systems due to a balance between recognition and performance [22]. CNNs can also be applied for the dynamic recognition of signs. Unlike the traditional 2D-CNN networks, 3D-CNNs can represent spatial and short-term temporal information together. Singh et al., for

example, applied a 3D-CNN for the dynamic recognition of ISL signs in 2400 videos displaying 20 gestures with the accuracy of 88.24% [19]. However, classic CNNs serve mainly as spatial models and lack the ability to capture long-distance temporal dependencies. Hence, the good results obtained by some studies on CNNs cannot be treated as the evidence of their ability for the reliable continuous-sign recognition [10],[11].

The initial development of automatic spatial feature learning was very much dependent on CNNs, with the pretrained architectures providing a useful way of boosting performance without the need to create all visual features from scratch. However, despite their advantages, classic CNN configurations remain restricted in terms of the handling of lengthy temporal dynamics, as well as the combining of different data streams. Because of these limitations, scientists have resorted to employing recurrent structures, attention mechanisms and different Transformer architectures [3],[4],[12],[16].

2.5 Temporal and Sequence-Based Approaches

The shift from static to dynamic sign recognition brought in a main requirement for the models on how to learn the type of a sign as it develops in time. The dynamic signing not only relies on the position of hands but also the sequence of movements, their direction, speed and the timing of motions. Based on our review, we have noticed that 3D-CNNs, LSTMs, Bi-LSTMs, temporal CNNs and attention-based networks are being increasingly used to overcome the problem [19],[20],[21].

2.5.1 Spatiotemporal and Recurrent Models

Using a 3D-CNN, Singh et al. reported the extension of CNN-based recognition techniques to the temporal domain that resulted in the recognition of 20 dynamic ISL gestures. The data was collected from 10 participants yielding 2,400 video samples with 88.24% accuracy reported by the authors [19]. 3D-CNNs are different than the 2D ones in a way that, 3D-CNNs process data in a way that both spatial and temporal information can be taken into account and spatiotemporal patterns can be learnt jointly. However, in spite of its pros, this method is computationally more challenging. Another solution is offered by recurrent architectures including mostly LSTMs and Bi-LSTMs, which allow for retaining information from the earlier data. Shetty et al. researched applications of PoseNet, HRNet and MediaPipe representations together with temporal models using INCLUDE data set that included 4,292 videos and 263 signs [18],[39].

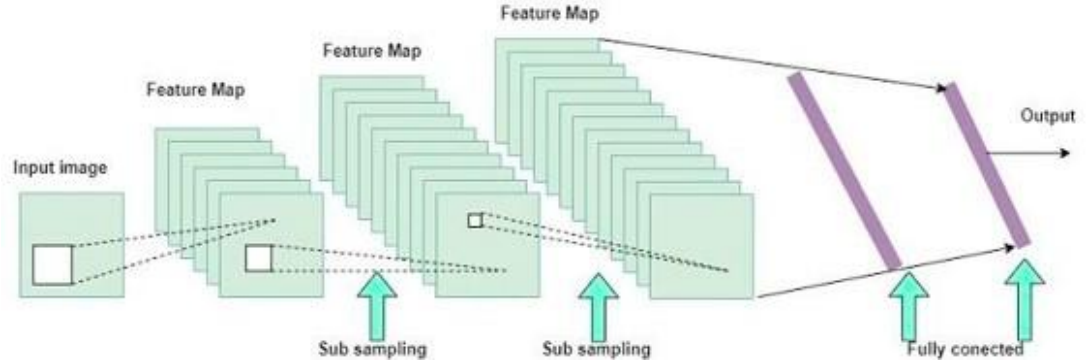


Fig. 5. General CNN architecture: convolution and sub-sampling layers extract feature maps from the input image, which are then classified via fully connected layers [37].

Their findings proved the significance of representation quality and found 98% accuracy for the MediaPipe-based technique used in the research. The use of convolutional operations makes it possible to avoid handling sequences in a recurrent manner. For example, Sa et al. fused elements of spatial and temporal convolutional networks and achieved 90.87% and 89.46% of accuracy in the context of CSL and ArSL datasets respectively [37]. Though the datasets do not belong to ISL, the work shows relevant methodological contribution to spatiotemporal modelling. Other research efforts, such as hierarchical Bi-LSTM with attention mechanisms, present other attempts of improving sequence modelling tasks [34].

2.5.2 Limitations and Transition Towards Attention

According to our observations, the temporal awareness does not depend solely on the sequence model but also on the quality and richness of the used representation. Some of the issues at play include the variability in the length of the sequences, absence of certain observations, presence of signer-dependent movement patterns, as well as computational complexity [10],[11]. It should be noted that even the recurrent models grapple with the task of making efficient use of connections between distant parts within long sequences due to the structure of their sequential processing. As a consequence, studies have increasingly proceeded in the sense of temporal attention and Transformer architectures, which have the ability to directly represent relationships over sequences [3],[4],[12]. Landmark and pose representations are being utilized alongside temporal representations as a means of reducing input dimensionality and focusing on motion information [5],[18]. The shift represents a transition from frame-based classification to improved understanding of multi-sequence data, ultimately paving the way for multimodal and transformer-based approaches.

2.6 Multimodal Approaches

Recognizing sign language is dependent on a variety of pieces of information that come together to help one identify the sign accurately. Among other things, hand size and hand movements give the most important information regarding the sign itself, while the movements of the body contribute to capturing exactly how the sign looks and is performed. Also, facial expressions, movements of lips and duration may be very useful during the process of recognizing the sign itself, especially since in some cases the sign is not simple and continuous [10],[11]. The trend is to gradually shift from approaches that are based on only one kind of information to multimodal systems that integrate data from RGB modalities and body movement tracking system [12],[24]. RGB information maintains the overall look of the signer while landmarks and pose provide a more organized description of hand and body movement. In this respect, using the two data modalities together allows for better understanding of sign language than if only one of them would be employed. Hence, the concept of multimodal fusion has been researched over the last couple of years for its potential improvement in signs' recognition [3],[4].

2.6.1 Motivation and Multimodal Representations

A major motivation for multimodal recognition is the limitation of relying exclusively on RGB information. Although RGB preserves rich visual and contextual information, it can also be affected by background, clothing, illumination and signer appearance [10],[11]. In contrast, hand landmarks and body pose provide more structured representations of configuration and movement [5],[18]. These modalities can therefore be viewed as complementary rather than competing sources of information.

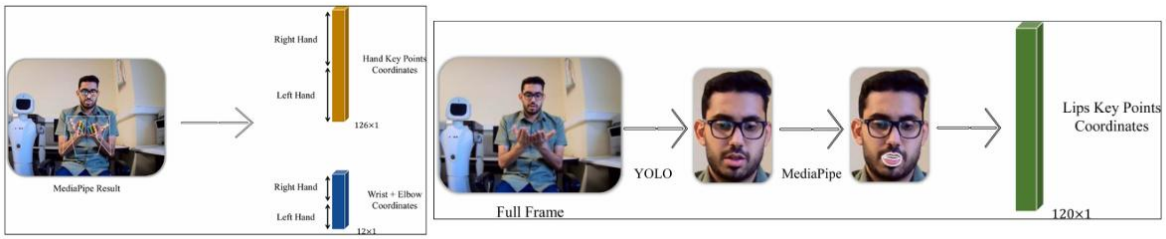


Fig. 6. Example of multimodal feature extraction from video frames: (a) hand keypoints with wrist and elbow coordinates, and (b) lip keypoint extraction using YOLO and MediaPipe [3].

Early multimodal approaches included wearable-sensor fusion. Studies combining sEMG, IMU and data-glove information demonstrated that different physical measurements could provide complementary descriptions of hand movement [14],[24]. However, the specialized hardware introduces additional requirements that can limit convenience and scalability compared with camera-based systems.Recent research increasingly focuses on vision-based multimodal fusion. Alkhoraif et al. proposed a two-stream appearance–pose architecture using Swin Transformer backbones and evaluated it on WLASL, MSASL and ASL Citizen, reporting top-1 accuracies

of 93.51%, 92.56% and 86.53%, respectively [12]. Although these datasets are not ISL datasets, the study provides relevant evidence for appearance–pose integration. TransMoVIF provides another important example by combining RGB visual information with skeletal motion through cross-attention. Its motion representation incorporates trajectory characteristics such as displacement, velocity and curvature, with reported accuracies of 96.80% on INCLUDE-50 and 93.95% on SMILE-DSGS [4].

2.6.2 Fusion Strategies and Challenges

Multimodal systems can combine information through early fusion, where modalities are combined before major processing; intermediate fusion, where modality-specific representations are jointly processed; late fusion, where predictions are combined; and cross-attention fusion, where one modality explicitly attends to another [3],[4],[12]. These strategies differ in complexity and their ability to learn relationships between modalities. Cross-attention is particularly promising because it can explicitly model complementary information between visual appearance and movement [4]. Ghadami et al. further demonstrated multi-stream fusion using hand, finger, lip, geometric and body-related keypoints, achieving 90.2% top-1 accuracy on ISLR101 [3]. Since this study concerns Iranian Sign Language, rather than Indian Sign Language, it should be interpreted as supporting methodological evidence for multimodal fusion rather than direct ISL evidence. Even though multimodal approaches offer better representations, they also create problems in areas of modality alignment, data loss, landmark detection mistakes and growing complexity of architecture another [3],[4],[12]. Poor performance in landmark extraction at the beginning of process can have negative consequences. Solving this task requires using architectures that can provide complementary features of different modalities without compromising efficiency. While working with ISL, utilizing visual features of practitioners and established landmarks is regarded as a way to improve effectiveness of signer-independent recognition systems. However, literature review shows that little research has been done in this regard, thus justifying the need for further research of Transformer architectures that can combine different modalities over time [5],[18].

2.7 Transformer-Based Approaches

Recent developments in the field of sign language recognition have been marked by the invention of the Transformer architecture. Transformer architecture differs from recurrent models in processing sequences: while conventional recurrent models deal with one element in the sequence at a time, Transformers employ a mechanism of attention to learn how different sequence parts relate to each other. Because of this, any sign language representation can benefit from using Transformer architecture as it is characterized by certain positional dependencies meaning that gestures in different parts of the sequence contribute to recognition of a sign language unit another [3],[4],[12]. The review reveals that the use of Transformer-based models for sign language recognition focuses not only on visual representation, but also on temporal modelling, multi-modal fusion and sign-to-text conversion [2],[26],[30]. It is beneficial to highlight that the newest architectures prefer to utilize Transformer architecture for extracting structured data, using both spatial and visual features another [3],[4],[12]. Nevertheless, the efficiency of this approach might be depended on the specifics of the data set, the challenges posed by the task, the choice of input representation and the experimental settings but computational limitations may also hinder the practical application of the technology [10],[12].

2.7.1 Visual, Temporal and Multimodal Transformers

One direction involves implementing ViTs for the purpose of visual sign detection. Ashfaq et al. introduced SignViT to classify uses of image patching and positional embeddings along with discrimination capability, ensuring a 99% success rate for the three investigated datasets, namely Urdu, Arabic and American Sign Language [30]. Although the work concentrates primarily on static detection and excludes Indian Sign Language, it still provides an example of the effectiveness of attention-based systems in visual sign description. A system of more relevance is TransMoVIF which combines methods of dual transformation of RGB vision features and skeletal movement separately, then merging them through cross-attention [4]. The motion pipe conveys motion characteristics, while the RGB path recognizes the raw RGB information. The reported accuracies of 96.80% on INCLUDE-50 and 93.95% on SMILE-DSGS demonstrate the potential of cross-modal attention for isolated sign recognition [4]. Transformer architectures are also utilized in activities involving

transitioning continuous sign-to-language. For instance, Yerimbetova et al. have used keypoints derived from MediaPipe and combined them with a TCN encoder and the Transformer decoder, using 225 keypoints in each frame, achieving around 75% accuracy on the KRSL dataset [27]. In addition, Ghadami et al. have applied a multi-stream Transformer-based approach that takes into consideration hand, fingers, lip movements, geometry and body features, achieving 90.2% accuracy on the ISLR101 dataset [3]. The two examples show how transformer architectures can provide a framework that allows combining multiple structured representations, even though the studies were conducted using different sign languages and settings.

2.7.2 Critical Assessment and Research Direction

According to our analysis, sign recognition using Transformer technology can be divided into three sections, namely Vision Transformers that are used for visual learning, temporal Transformers which are applied to sequences and finally multimodal Transformers which deal with interactions among different representations [3],[4],[12]. However, these categories should not be considered similar. The type of Transformer which is used for pictures tackles a different kind of challenge compared to its multimodal counterpart which works with RGB and skeletal sequences. The main benefit of Transformers is the attention mechanism that allows to model the connections between various components of representation or sequence in a more straightforward way [3],[4],[12]. This is especially true for sign languages in which comprehension may rely on gestures happening at various moments of time. At the same time, systems using Transformer technology bring obstacles regarding processing requirements, preprocessing pipeline, fusion complications and reliance on properly representative information for training [10],[12]. A complex structure cannot replace limited vocabulary, signer spread, or poorly represented data. Our review indicates that there are still numerous recent Transformer papers which analyze only isolated recognizers or non-ISL data instead of focusing on ISL recognition that is continuous and signer-independent [27],[30]. This means that improvements mentioned should be viewed concerning the data and tasks which were used instead of assuming that the results achieved by Transformers are better than the results achieved by CNNs or recurrent models in all cases. The results of using Transformers should be analyzed more in the sense of their ability to build long-range correlations rather than to substitute previous models. Because of this, multimodal Transformers could be important for making more effective and general ISL recognition systems [3],[4],[12].

2.8 Real-Time and Lightweight Recognition

In order to be practicable for real-life communication, a recognition system must not only achieve a high level of accuracy in recognition. The performance of a certain model on standard datasets may be good, but the model may still face problems in application if the feature extraction process is slow, time-consuming or requires heavy hardware. Therefore, the researchers have focused on lightweight models, fast feature extraction and alternative methods of work that are economically saved regarding the use of computational resources [6],[8],[9],[22].

2.8.1 Lightweight and Efficient Architectures

A number of research works have been conducted to find the ways of lowering computation by achieving good recognition performance. Raja et al investigated the use of YOLO-based approaches like YOLOv5-YOLO11 and YOLO-NAS in real-time ISL detection based on the Image-ISL-PICT dataset. YOLO-NAS was able to gain 0.983 mAP@0.50 with latency below milliseconds which shows that object-detection models can be highly efficient for ISL detection [9]. According to Gogoi et al the accuracy for ISL detection stood at 99.64% when lightweight CNN was used with latency of 76.38 ms [7]. The comparison of VGG16, VGG19, ResNet50 and MobileNetV2 proves the importance of lightweight architectures in practice [6],[8]. Other works have used knowledge distillation, spiking neural networks and other models including MobileNetV3, MobileViT and ShuffleNetV2 in order to reduce the complexity of models but keep the recognition ability [22],[38].The findings demonstrate that effectiveness is increasing in importance in the field of ISL and related sign language recognition research. However, numerous light-weight methods continue to concentrate on static or isolated signs [6],[7],[8]. Efficient image-based as well as fast detection cannot incorporate the temporal information necessary for continuous signing [18],[19].

2.8.2 Deployment Considerations

The speed of the final classifier is only one element of real-time recognition. A more comprehensive approach requires additional steps: video acquisition, hand detection, feature extraction and temporal computation followed by output generation. Each of these steps has its time-consuming aspects which may adversely affect the overall operation duration, particularly under multi-stage processing conditions [5],[9]. A critical factor is the continuous signing, whereby the system needs to process incoming frames without delays. It is relevant to highlight the dichotomy between efficiency and model capacity to absorb information: lightweight CNN and landmark-based methods require little effort to be employed, while multimodal and transformer-based approaches demand more computational power to get more complex information [4],[12]. However, it is not always easy to reduce the complexity of the model since it may lose its ability to recognize complicated signing patterns, while complex methods may be of no help for resource-constrained devices. The literature available suggests that practical ISL recognition must take into account these aspects such as accuracy, temporal capability, robustness and computational efficiency [9],[10]. Therefore, future studies should focus on effective representations of relevant features and the development of light architectures and methods for model compression while preserving spatial and temporal data vital for accurate continuous sign recognition.

3. Comparative Analysis of Datasets, Modality, Architectures and Performance

3.1 Comparative Analysis of Datasets

Upon analysis of the datasets utilized in the 36 studies reviewed, it is noted that characteristics of the datasets play an important role in interpreting reported recognition performance. The studies vary significantly with respect to the language, size of the sample, size of the vocabulary, modality, signer population, conditions of acquisition and task definition. Hence, when interpreting reported performance, it is necessary to take into consideration both the characteristics of the dataset and the architecture of the model [10],[11].

3.1.1 Size, Vocabulary and Task Complexity

Dataset size by itself does not reflect the quality of a dataset as well as the difficulty of recognition. The researchers Katoch et al. make use of 36,000 images within 36 classes and 3 signers whereas Singh and colleagues use 2,400 videos representing 20 dynamic gestures in their work [13],[19]. The INCLUDE dataset introduces a different dimension, with 4,292 videos covering 263 signs, while Image-ISL-PICT dataset consists of 3,580 images belonging to 115 classes and 10 signers [9],[39]. All these examples illustrate that sample size, vocabulary size and temporal complexity are different characteristics of difficulty in regard to datasets. A large static dataset and a smaller video dataset cannot be considered equivalent simply on the basis of sample count.

3.1.2 Static vs. Dynamic and Signer Diversity

A crucial difference between static image datasets and moving video datasets exists. Static datasets are mostly used for letter, numeral and isolated-letter detection and usually achieve very high performance in highly controlled environments [6],[8],[13]. Dynamic datasets keep longitudinal data and require methods to learn transitions, movements and relationships [18],[19],[20]. A reported accuracy of 99% on a static alphabet dataset should not be interpreted as equivalent evidence of continuous-sign recognition. Another factor influencing generalization is variability among referents. Variability in hand shape, movement, signing speed, body position and other physical features can lead to considerable variability of the system performance [23],[34]. Many datasets examined in this study contain rather small groups of referents, making it impossible to see how the models would generalize to novel individuals. Signer-independent evaluation is therefore an important consideration when interpreting reported results [12],[34].

3.1.3 Custom vs. Benchmark Datasets

Researchers favour custom datasets for their collection of particular lexicons along with full control over experimental settings. However, variations in the recording environment, camera types, backgrounds, annotation methods, signer participants and train-test split procedure render comparisons complicated [10],[11]. Standard resources in research such as INCLUDE, INCLUDE-50, WLASL, MSASL, ASL Citizen as well as ISLR101

offer a more stable benchmark despite the fact that some of them refer to the languages that differ from ISL and should be thus regarded as methodological examples [39],[42],[43],[44].

3.1.4 Representation and Evaluation Implications

The latest studies show that recent progress has gone beyond relying on heavy pixel-level RGB data in favour of structured spatial positioning. Indeed, by compressing input data to the critical information regarding hand and body landmarks, it is possible to lower input dimensionality, while preserving the spatial and kinetic structure [5],[18]. The problem is that any mistakes made while extracting the landmarks in the first phases might continue appearing at the later stage of temporal modelling and impact the recognition efficiency [5],[18]. In this context, one can say that an in-depth analysis of these studies shows that the complexity of the dataset is more important than its raw size. The essential factors in this case involve vocabulary size, diversity of signers, number of variations in time, conditions of acquisition, quality of annotation and assessment methods [10],[11]. This indicates that the current technological industry should develop standardized British ISL testing environment according to the principle of creating a continual feed of videos and using a signer independent measurement strategy. Such resources would enable more meaningful comparison of recognition systems and provide stronger evidence of real-world generalization.

3.2 Modality Comparison

Input modalities are continuously changing from isolated sensors to structured multi-stream systems. Typical RGB video works well with regular cameras but can be influenced by background, lighting, outfit and the appearance of the signer. Therefore, researchers have started to adopt hand landmarks and skeleton trackers as tools to convert inputs into structured coordinate data [5],[18]. On the other side of the spectrum, devices like IMUs and sEMG arrays together with data gloves supply direct kinematic information, but at the price of additional equipment and less natural way of interaction [20],[21],[24]. Using only one data source is often insufficient to capture the whole variety of features of sign language. That is why it became a common practice in modern architectures to combine RGB video with skeleton path, hand coordinates and other non-manual cues in order to use their advantages [3],[4],[12]. Therefore, the goal of researchers has changed: instead of trying to find the “best” kind of input they are increasingly working on the development of models that integrate various types of data successfully.

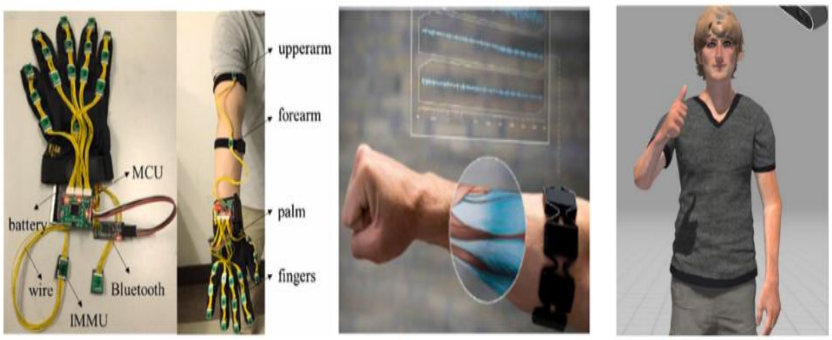


Fig. 7. Different approaches of hand-information acquisition approaches: (a) wearable data-glove sensors, (b) EMG-based wearable sensing and (c) vision-based sensing using cameras [11].

Current research indicates that RGB video, hand landmarks and body poses can be used in combination for sign language recognition purposes. Using these three types of data together allows for improvement in visual presentation, hand shape and body motion recognition while also encouraging the exploration of efficient multimodal fusion techniques and Transformer-based configurations [3],[12],[18].

Table 4. Comparison of major input modalities used in ISL/SLR, highlighting the information captured, strengths, limitations and typical applications.

Input modality	Information captured	Major strengths	Major limitations	Typical use in reviewed studies
RGB images	Appearance, hand shape, visual context	Rich information; camera-based; no special hardware	Sensitive to background, lighting, clothing and appearance variation	Static and isolated sign recognition; CNN/ViT
RGB video	Appearance + temporal movement	Preserves spatial and temporal information	Higher computational and storage requirements	Dynamic/continuous recognition
Hand landmarks	Hand joints and configuration	Compact, structured, lower-dimensional representation	Dependent on landmark detection; may lose appearance information	MediaPipe + LSTM/TCN/ML models
Body pose / skeleton	Body joints and movement	Captures body-level spatial and motion information; less dependent on appearance	May miss fine-grained hand details	Pose-based and multimodal recognition
Wearable sensors	Physical movement/kinematics	Direct and structured motion measurement	Requires specialized hardware; less natural interaction	Data glove, IMU, sEMG systems
RGB + pose/skeleton	Appearance + body/skeletal motion	Combines complementary visual and structural information	More complex preprocessing and fusion	Multimodal Transformer systems
Hand + body + other keypoints	Fine hand + body + non-manual cues	Richer representation of signing	Higher feature and computational complexity	Multi-stream Transformer approaches
Multimodal combinations	Multiple complementary representations	Potentially greater robustness and information coverage	Synchronization, missing modalities and computational cost	Recent multimodal research

3.3 Comparative Analysis of Model Architectures

The analysis of our 36 studies shows that there has been a transition from traditional classifiers based on certain features to the use of deep learning techniques and techniques providing for temporality, multimodality and using the Transformer architecture. SVM, KNN and MLP were among the first to provide an indication of classifiers some time ago, while CNN provided for spatial learning of features automatically [13],[16],[17]. After that, 3D CNN, LSTM and temporal CNN allowed applying the recognition results to moving sequences, while landmark- and pose-based models created some more sophisticated forms of representation of the situation [19],[20],[21]. Modern approaches based on deep learning and Transformers are out mostly capable of considering temporal dependences and connections between logics [3],[4],[37]. We also observed that more complicated architectures do not mean better identification. The performance described depends on many aspects such as type of the dataset, vocabulary, definition of the task, modality and method of assessment

[10],[11]. For instance, while CNN demonstrates great results on a limited static dataset, Transformer can work with a more complicated tasks involving video and different modalities [6],[8].

Thus, to ensure a proper comparison, we should consider the target problem and task complexity more than achieve maximally high scores in benchmarks. Moreover, technological development means that new deep learning techniques are created to overcome weaknesses of the previous ones, starting from learning spatial representations and finishing with temporal modelling and multimodal integration [19],[20],[21]. In the future, it is quite improbable that simple increase of the model's size brings substantial advancements. Instead, new architecture will need to balance between recognition accuracy, modal integration, temporal awareness and low latency [9],[12]. Table 5 summarizes these model families along with their capabilities, strengths, limitations and typical tasks.

Table 5. Comparison of major model architectures used in ISL/SLR recognition, highlighting their primary capabilities, strengths, limitations and typical tasks.

Model family	Primary capability	Strengths	Main limitations	Typical task
Classical ML (SVM, KNN, MLP)	Feature-based classification	Simple, relatively low computational cost; effective with engineered features	Dependent on handcrafted representations; limited feature learning	Static/isolated recognition
CNN	Spatial feature learning	Strong automatic visual representation; effective for images	Limited inherent temporal modelling	Static/isolated signs
Transfer learning	Reuse of pretrained visual features	Useful when task-specific data are limited; strong performance	Pretrained features may not fully capture sign-specific characteristics	Static/isolated recognition
3D-CNN	Spatial + short-term temporal learning	Processes spatial and temporal dimensions jointly	Computationally expensive; limited long-range modelling	Dynamic video recognition
LSTM/Bi-LSTM	Sequential/temporal modelling	Captures temporal dependencies; suitable for sequences	Sequential processing; difficulty with very long-range dependencies	Dynamic/continuous recognition
Temporal CNN/TCN	Temporal pattern learning	Parallel temporal processing; effective sequence modelling	Still dependent on quality of input representation	Dynamic recognition
Landmark/Pose + sequence model	Structured motion learning	Compact representation; reduces dependence on raw appearance	Sensitive to landmark/pose detection errors	Dynamic/isolated recognition
Multimodal models	Cross-modal representation	Combines complementary information	Increased preprocessing and computational complexity	Multimodal recognition

Vision Transformer	Global visual relationships	Attention-based global feature modelling	Higher data/computational requirements	Image/visual recognition
Multimodal Transformer	Temporal + cross-modal relationships	Long-range dependency modelling and multimodal fusion	Computationally demanding; fusion design is complex	Dynamic/multimodal recognition
Lightweight models	Efficient inference	Lower latency and resource requirements	May trade some representational capacity for efficiency	Real-time recognition

3.5 Comparative Analysis of Performance

The diverse nature of the literature in question does not make it easy to carry out a direct empirical comparison. The different metrics employed in the studies examined cover a large range from the widely-used classification accuracy and F1-scores to the more sophisticated localization mAP and processing throughput. Moreover, they are employed under different environmental conditions, input modalities and validation protocols [10],[11]. Therefore, it would be misleading to even attempt to form a proper performance hierarchy for the mentioned metrics. Each different metric is designed to address a specific aspect of the system, which may include gesture classification, object recognition, system performance and language generation; therefore, their findings should be interpreted in relation to the specific task, dataset and architecture used [9],[32],[37].

Table 6. Comparison of commonly used evaluation metrics in ISL/SLR research, including their definitions, advantages, limitations and appropriate use cases.

Metric	Formula	Pros	Cons	When to Use
Accuracy	(TP+TN) / (TP+TN+FP+FN)	Simple, intuitive, easy to communicate	Misleading on imbalanced datasets	Balanced class distributions; quick baseline comparison
Val. Accuracy	Accuracy on held-out validation set	Tracks generalization; detects overfitting	Still misleading on imbalanced val sets	Model selection, early stopping, hyper parameter tuning
Precision	TP / (TP + FP)	Penalizes false positives strongly	Ignores false negatives	When false positives are costly (e.g., spam detection)
Recall	TP / (TP + FN)	Captures all actual positives	Ignores false positives	When false negatives are critical (e.g., disease detection)
F1-Score	2 × (P×R) / (P+R)	Balances precision and recall	Treats P and R as equally important	Imbalanced classes; when both FP and FN matter
mAP	mean of AP across classes / IoU thresholds	Evaluates localization + classification together	Computationally heavy; hard to interpret alone	Object detection benchmarks (COCO, PASCAL VOC)
BLEU(Generation)	BP × exp(Σ w_n log p_n)	Fast, reference-based, language-agnostic	Poor semantic understanding; penalizes valid	Machine translation, text summarization

			paraphrases	
ROUGE(Generation)	Recall of n-gram overlap vs reference (ROUGE-N, ROUGE-L)	Recall-oriented; ROUGE-L captures word order	Surface-level; misses semantic similarity	Summarization, question answering evaluation
Loss	e.g., $-\Sigma\, y_i \log(\hat{y}_i)$ (cross-entropy)	Directly optimized; shows learning progress	Not interpretable to non-ML audiences	Monitoring training & validation curves
Validation Loss	Same as loss, on held-out val set	Detects overfitting early	Depends heavily on val set quality	Hyper parameter tuning, early stopping decisions
Latency	t_output − t_input (ms or s)	Directly reflects real-world speed	Varies with hardware, batch size, load	Real-time systems; deployment & edge applications
FPS	Frames processed / second	Intuitive speed measure for vision tasks	Hardware-dependent; not reproducible across setups	Video processing, real-time object detection pipelines

Methodological evaluation in the literature is very controversial and relates in detail to the type of recognition task. While classification accuracy, precision and F1-scores prevail in basic sign recognition, translation tasks call for evaluations based on text generation including BLEU and ROUGE scores [32]. Once the task has evolved into spatial detection or real-time usage, the measures further develop and focus on mAP, processing latency and throughput (FPS) [9]. In a nutshell, this indicates that any model may be evaluated properly when one obtains understanding of both metric selected and the underlying difficulty of the data and the task architecture employed by it [10],[11].

Table 7. Representative performance of ISL/SLR approaches across task categories, with reported results and their corresponding implications.

Study / approach	Language	Task	Dataset	Reported result	What the result indicates
Sharma & Anand [17]	ISL	Static alphanumeric	ISL dataset	99.76% numerals; 97.0% alphabets	Strong static spatial recognition
Katoch et al. [13]	ISL	Static recognition	36,000 images; 36 classes	99.64% CNN	Strong controlled image classification
Ingoley& Bakal [8]	ISL	Static recognition	Custom ISL	>99% validation for several models	Transfer learning can perform strongly on static tasks
Singh et al. [19]	ISL	Dynamic recognition	2,400 videos; 20 gestures	88.24%	Temporal recognition is more challenging
Shetty et al. [18]	ISL	Dynamic gesture-to-text	INCLUDE	98% on reported 43-class MediaPipe experiment	Structured temporal representation can be effective
Ghadami	Iranian	Isolated SLR	ISLR101	90.2% top-1;	Multistream

et al. [3]				93.1% top-5	Transformer fusion
Mishra et al. [36]	ISL	Gesture-to-sentence	Images + video/keypoints	99.98% letters/numbers; ~87% sentence	Language-level task is noticeably harder
Rajet al. [9]	ISL	Real-time detection	Image-ISL-PICT	0.982 mAP@0.50	Strong detection performance, but static
Gogoi et al. [7]	ISL/ASL	Gesture-to-speech	Public datasets	99.64% ISL; 13.62 ms minimum inference	Accuracy and efficiency can be combined in constrained tasks

Studies in static ISL often report very high rates of accuracy with numbers provided by Sharma and Anand reaching 99.0% and by Katoch et al. from CNN coming to the level of 99.64%. On the other hand, a research study conducted by Singh et al. reported an accuracy of 88.24% with 3D-CNN in dynamic recognition while multimodal methods and Transformers such as TransMoVIF and Ghadami et al. managed to achieve accuracy rates of 96.80% and 90.20% respectively in their experiments [3],[4],[19]. This indicates that success in numerical parameters can't be analyzed separately from level of complexity of the task. Instead of depending on single indicators, we propose four essential dimensions for evaluation purposes. Recognition quality serves as the primary dimension and includes conventional accuracy measures, F1-score, or mAP depending on the problem being solved. This is compared with task complexity which breaks down data structures into static, isolated or variable formats. Another important factor is practical relevance that requires generalization performance, namely, signer-independence and condition-independent capabilities, as well as efficiency parameters, such as size, dimensions and inference time [9],[10],[11].

A classifier that has a 99% accuracy rate works unlike any ML system with a lower score, for example a 90% scoring multimodal classifier and in these cases, making a small jump upward in accuracy means so much more when it is done in a more complex real-world setting. Foundational studies have shown strong performance of classifiers trained on isolated gestures but these results cannot be extrapolated to new architectures in other settings [17],[19]. Thus, progress in the field must be assessed according to the structure of data, complexity of tasks and requirements for underlying hardware. In general, the real frontier in ISL study is the ability to deliver reliable recognition results for systems moving to dynamic conditions and to user-independent environments [5],[18].

3.6 Common Limitations Identified Across the Literature

3.6.1 Dataset and Generalization Limitations

The diversity and standardization of datasets is a serious problem that causes much trouble. Most of the studies run on their unique datasets which have different vocabulary, signers, conditions and annotation methods making the comparison hard [10],[11]. It is also important to note that the high number of samples does not mean the high diversity of the dataset. In other words, thousands of samples from a few signers do not bring much evidence for generalization. The importance of signer-independent evaluation is stipulated by the fact that each signer has his/her special hand shape, movements, signing speed and looks [23],[34].

3.6.2 Static Recognition and Temporal Challenges

The majority of high precision benchmarks focus on static alphabets, individual gestures, or digits [6],[8],[13],[17]. Although these limited problems can provide helpful benchmarks, they do not accurately reproduce the temporal properties of continuous sign language. Indeed, proper conversation recognition demands architectures capable of handling the spatial-temporal properties, variable speeds, sequence boundaries and context dependencies involved in this sign language [20],[21]. Hence, one should not equate high static precision with strong continuous recognition.

3.6.3 Representation and Modality Limitations

Different types of input modalities come with their own limitations. For example, RGB images provide great visual detail but are sensitive to changes in background, lighting and appearance of objects. On the other hand, watching hand landmarks and pose offers compact structural information but requires reliable keypoint detection systems, which may lead to loss of some visual information [5],[18]. While wearable sensors allow for capturing structured motion information, they require specific hardware [14],[24]. In addition, multimodal approaches are capable to overcome some of the limitations of these modalities but bring their own challenges in terms of synchronization, preprocessing, missing information and complexity of fusion [3],[4],[12]. Using different modalities together may help researchers overcome some of these difficulties, although some technical difficulties arise as well.

3.6.4 Performance and Deployment Challenges

Due to differences in method, terminology and evaluation methods, comparisons between systems are generally quite inconsistent. Trying to determine how accurately a model is working without placing the experimental conditions in the context could be risky [10],[11]. Thus, there is an obvious trade-off between execution depth and runtime efficiency. Bulky transformer networks and multi-stream fusion methods are able to generate complex representations but are more computationally demanding [3],[4],[12]. Conversely, lightweight solutions, such as MobileNet, YOLO-based approaches and light versions of neural networks, focus on speed of performance [7],[8],[22]. As shown in the previously mentioned literature, the ability to apply ISL technology successfully will rely on a balance between the accuracy of the gesture recognition system, continuous sequencing, meeting the needs of the end-user and keeping the technology cost-effective [5],[7],[9].

3.7 Critical Findings from the Comparative Analysis

The comparative study of the 36 examined studies shows that there are a number of trends that can be observed in the research on ISL and sign language recognition. The development of the field has moved from the use of handcrafted features and classical classifiers to the use of more sophisticated solutions, such as CNN, temporal approaches, body pose, hand landmarks, multimodal systems or Transformers [13],[19],[20],[21]. The results obtained should be always interpreted considering the experimental context. High accuracy results can be obtained in static and limited tasks, while recognition in dynamic context involves much more temporal and contextual complexity [18],[37]. The usage of temporal and multimodal techniques confirms this tendency that RGB allows to gather visual and contextual information, while pose and landmarks provide more structured representations [4],[5],[18]. However, when using multimodal methods, new problems arise, such as synchronizing information, missing information, cross-modal interaction and costs of computation [12].

The studies reviewed imply that Transformers are evolving into something more than a simple alternative choice to CNNs or recurrent architectures, owing to their attention mechanism that allows them to provide a model for long-distance time dependencies and relationships between different representations, including applications related to sequence-to-text conversion [3],[4],[30]. Even though this progress has been achieved, the practicality of the process is still not verified more than the theoretical measurements and benchmarks themselves, which is of particular concern when it comes to performance in real-world conditions that include unseen signers, changes in environment, occlusion, changes in speed of signing and struggles with landmark detection [10],[11]. Therefore, the key task now is to create systems that are capable of not only attaining the highest possible accuracy across classes but also combining different visual and structural features, accounting for time dependencies, generalizing across signers and environments and being computationally optimal [4],[5]. These findings provide the basis for the research gaps discussed in the following section.

4. Research Gaps

When we examined and compared the 36 reviewed studies, we identified several interconnected gaps in ISL/SLR research. Other than representing isolated limitations, these gaps indicate areas where current approaches remain fragmented, insufficiently evaluated, or difficult to translate into practical recognition systems. The literature demonstrates substantial progress in individual components such as spatial feature

extraction, temporal modelling, landmark-based representations and Transformer architectures; however, these developments have not yet been consistently integrated into a unified and robust ISL recognition framework [10],[11].

4.1 Limited Unified Multimodal Integration

Despite the increasing number of studies that uses multimodality, there has been little effort made toward unified approaches to ISL recognition that are based on RGB, hand landmarks and body pose. Current research studies focus on different combinations of modalities, including RGB with motion, pose features or landmark-based representations independent of raw images [4],[5],[12]. This is a chance to explore the functionalities of the modalities in integration. RGB means rich visual information, while hand sections and body poses add the structure to the data while handling space and movement of the sign itself. Consequently, multimodality of the recognition will not hurt the understanding of the signing as compared to visual means. However, multimodal recognition is not as easy as mixing the elements because there are differences in the modes being represented, time, reliability of the mode and interaction between modalities. This remains an important gap in current ISL research [3],[12].

4.2 Limited Continuous and Context-Aware Recognition

While much research has been performed on static and isolated signs [13],[19], continuous recognition has received considerably less attention. Continuous signing presents challenges that do not arise in isolated classification. In continuous recognition, researchers must address transitions between signs, variable sequence lengths, the determination of temporal boundaries and the relationships between successive signs. Models must therefore recognize not only the position of each gesture but also the surrounding context and how it changes over time [20],[21]. This distinction is critical for practical communication, since signs are rarely produced with clear separation in natural use. As a result, a system that performs well on isolated signs may still struggle with overlapping signs, fast transitions, or unusual sign sequences. This gap compounds the integration challenge raised in Section 4.1, since a genuinely unified model must handle continuous input rather than pre-segmented isolated signs. Greater effort is therefore needed on continuous and context-aware ISL recognition, supported by more adequate evaluation methods and datasets that better reflect natural signing [10],[11].

4.3 Limited Signer-Independent and Standardized Evaluation

This review shows considerable variety among the databases, including differences in signer characteristics, recording conditions, environmental contexts, lighting and camera parameters. In addition, differences in data types, task definitions and training/testing setups make it difficult to compare results across studies [10],[11]. This is particularly true for databases with a limited number of signers, which restricts the ability to draw reliable conclusions about a model's generalization [18],[34]. Thus, high accuracy alone is not sufficient evidence that a model generalizes well. A model may perform well when training and testing conditions are similar, but its performance can decline when the signer or recording environment changes. This limitation also affects the multimodal and continuous-recognition gaps discussed earlier, since the results reported in those sections cannot be reliably compared without evaluation protocols that are independent of signer characteristics.

4.4 Accuracy–Efficiency Trade-off

The use of multimodal and transformer-based architectures enables better representations as well as better modelling capabilities but at the same time requires more computational resources [3],[4],[12]. In contrast, light-weight architectures including a variety of mobile network-based architectures, efficient CNNs as well as optimized frameworks for detection serve better in terms of availability of resources and for real-time use [7],[9],[22]. This presents a substantial trade-off of accuracy and simplicity.

The challenge therefore consists of not only improving recognition accuracy but assessing architectural complexity in accordance with its applicability. A model that increases recognition accuracy in exchange for needing significantly greater computational resources is not likely to be more applicable than a slightly less accurate model that works on available hardware. Therefore, ISL systems should take into account parameters

such as accuracy, robustness, temporal capability, complexity of models and inference latency at the same time rather than maximizing any of these parameters separately [9],[10].

4.5 Central Research Gap

The analysis in Sections 4.1 to 4.4 brings out the single major limitation: although major advances have been made in terms of multimodal fusion, continuous recognition, standardized evaluation and computational efficiency as separate domains, no research so far has been able to put all these advances together in a single frame. This leads to the formulation of the principal gap revealed in this review: the absence of a unified multi-stream framework for obtaining complementary visual and structural information. In particular, such a multi-stream system would be required to use RGB frames, detect hand landmarks and skeleton coordinates, interpret their interconnections and recognize them, while doing all this in real time. Closing this gap would contribute to making ISL recognition move beyond different patches applied to particular fragments of the model into uniform and practical solution.

5. Conclusion And Future Directions

5.1 Conclusion

Our review covering 36 different studies confirms that there is an evident positive transition regarding ISL recognition from manually-created features and classical machine learning to CNN-based recognition, temporal models and landmark and post representations and multimodal architectures to transformers [19],[20],[21]. Each stage corrected shortcomings of previous technologies leading to improvements in spatial recognition, temporal analysis and collaboration of different information. As for the comparative evaluation, it also turned out that up-to-date studies have very good recognition results on static and isolated-sign recognition tasks while dynamic and continuous language tasks still require significant effort [18],[37]. The differences between datasets used in the studies, vocabulary, the diversity of signers, definitions of tasks and conditions of evaluation create obstacles for straightforward comparison of the findings obtained by studies [10],[11]. The proof strongly suggests that the primary concern is no longer just reaching a higher classification accuracy level. To be usable in practice, systems will have to tackle temporal comprehension, signer-independent generalization, multimodal integration, robustness and efficiency at the same time [5],[9],[12].

5.2 Future Directions

Looking at the gaps that were identified in this review for future ISL research, it is important that we must find larger, more diverse and standardized datasets comprising the larger vocabulary, different signer types, continuous video, efficient conditions of recording of the video and must examine procedure for annotation and evaluation protocols as pointed out in several studies [10],[11]. In particular, we must emphasize on continuous recognition, or context-aware recognition that can recognize changes of the different lengths of the sequences, signing transition and changes with respect to the temporal relationships, among others [21],[37]. Further, advancement in multimodal fusion must be concentrated by combining the RGB images, the positions of the hands and the movements of the body in combination with other relevant features like lip and face [3],[4]. Lastly, the improvement in the processing efficiency must be looked at concurrently with recognition performance by getting the lightweight structures and efficient implementation as argued in various studies [7],[9]. Conceptually, this direction can be seen in figure 8 that outlines a conceptual framework combining RGB, hand-landmark and body-pose inputs through multimodal fusion and Transformer-based temporal modelling to produce ISL output.

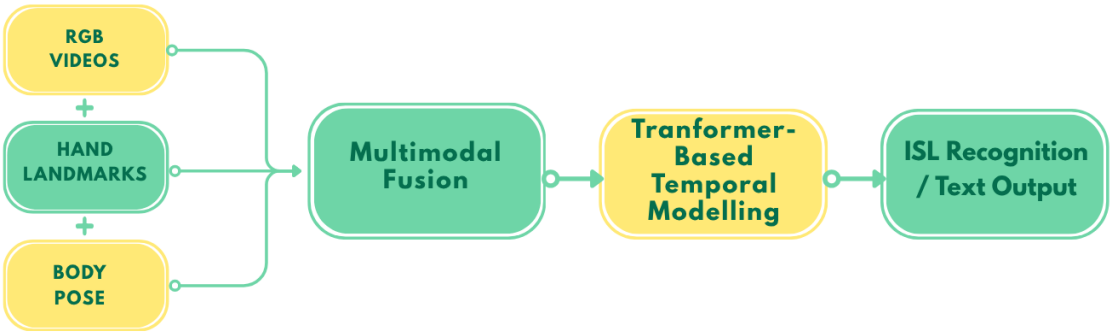


Fig. 8. Conceptual framework for multimodal ISL recognition, combining RGB, hand-landmark and body-pose representations via multimodal fusion and Transformer-based temporal modelling.

Emerging from this review as an especially promising possibility is the utilization of RGB video, hand attributes and body pose motion in the multimodal transformer structure. The RGB video allows to keep visual and contextual data, while the hand attributes allow to capture detailed configurations of the hand, in turn, the body pose can provide general structural motion. The transformer's attention serves in modelling temporal links and interrelations of the mentioned representations [3],[4],[5],[12]. At the same time, the use of multimodal transformers is not a guarantee that the solution is guaranteed. Rather, their efficiency greatly depends on the variability of the dataset, quality of features extraction, fusion approaches, techniques of temporal modelling and evaluation methods [10]. Hence, the analysis provided in this review shows the shift of ISL recognition from traditional gesture classification to the one that goes beyond mere classification and includes the understanding of signs that takes into account not only time factors, but also the use of different modalities for this purpose [10],[11].

References

[1] World Health Organization. (2026, March 3). *Deafness and hearing loss.*

[2] Woll, B. (2024, September 20). *What have sign languages taught us about human language development?* The British Academy Article.

[3] Ghadami, A., Meghdari, A. F., & Taheri, A. (2027). A transformer-based multi-stream approach for isolated Iranian sign language recognition. *Computer Speech & Language, 101*, 102017. https://doi.org/10.1016/j.csl.2026.102017

[4] Renjith, S., Varghese, A., Rashmi, M., & S. S., P. (2026). Transformer-based motion-visual integrated fusion for isolated sign language recognition. *Computers and Electrical Engineering, 130*, 110902. https://doi.org/10.1016/j.compeleceng.2025.110902

[5] Patel, R., Mehta, V., Patel, M., & Raval, D. (2026). A MediaPipe-powered model for static Indian sign language recognition. *Procedia Computer Science, 283*, 2220–2229. https://doi.org/10.1016/j.procs.2026.06.294

[6] Ingoley, S., & Bakal, J. (2025). Interpretation of Indian sign language to text and speech to communicate with speech and hearing-impaired community. *Procedia Computer Science, 258*, 1980–1992. https://doi.org/10.1016/j.procs.2025.04.449

[7] Gogoi, P., Karsh, B., Karsh, R. K., Laskar, R. H., & Bhuyan, M. K. (2025). Vision-based real-time gesture-to-speech translation for sign language. *Procedia Computer Science, 258*, 2050–2059. https://doi.org/10.1016/j.procs.2025.04.455

[8] Ingoley, S., & Bakal, J. (2025). Efficient models based on deep learning technique for Indian sign language. *Procedia Computer Science, 258*, 318–331. https://doi.org/10.1016/j.procs.2025.04.269

[9] Raja, R., Sreemathya, R., Turuka, M., Jagdale, J., & Anisha, M. (2025). Performance comparison of different versions of YOLO for Indian sign language captioning in real time of multiple signers. *Procedia Computer Science, 259*, 991–1000. https://doi.org/10.1016/j.procs.2025.04.053

[10] Trabelsi, L., Harrouch, H., Mohamed, S., Jebali, M., & Sharma, A. (2026). Advancements and challenges in vision-based sign language recognition: A comprehensive review. *Information Fusion, 126*, 103626. https://doi.org/10.1016/j.inffus.2025.103626

[11] Jalayer, R., Jalayer, M., Orsenigo, C., &Tomizuka, M. (2026). A review on deep learning for vision-based hand detection, hand segmentation and hand gesture recognition in human–robot interaction. *Robotics and Computer-Integrated Manufacturing, 97*, 103110. https://doi.org/10.1016/j.rcim.2025.103110

[12] Alkhoraif, A. A., Alsulaiman, M., Abdul, W., & Bencherif, M. (2026). Ensemble transformer-based word-level sign language recognition with multi-modal input fusion. *Journal of Engineering Research, 14*, 738–747. https://doi.org/10.1016/j.jer.2025.07.006

[13] Katoch, S., Singh, V., & Tiwary, U. S. (2022). Indian sign language recognition system using SURF with SVM and CNN. Array, 14, 100141. https://doi.org/10.1016/j.array.2022.100141

[14] Kakoty, N. M., & Sharma, M. D. (2018). *Recognition of sign language alphabets and numbers based on hand kinematics using a data glove*. Procedia Computer Science, 133, 55–62. https://doi.org/10.1016/j.procs.2018.07.008

[15] Tripathi, K., Baranwal, N., & Nandi, G. C. (2015). Continuous Indian sign language gesture recognition and sentence formation. Procedia Computer Science, 54, 523–531. https://doi.org/10.1016/j.procs.2015.06.060

[16] Krishnaveni, M., & Radha, V. (2012). *Classifier fusion based on Bayes aggregation method for Indian sign language datasets*. Procedia Engineering, 30, 1110–1118. https://doi.org/10.1016/j.proeng.2012.01.970

[17] Sharma, P., & Anand, R. S. (2021). *A comprehensive evaluation of deep models and optimizers for Indian sign language recognition*. Graphics and Visual Computing, 5, 200032. https://doi.org/10.1016/j.gvc.2021.200032

[18] Shetty, S., Hirani, E., Singh, A., & Koshy, R. (2024). Gesture-to-text: A real-time Indian sign language translator with pose estimation and LSTMs. *Procedia Computer Science, 235*, 2684–2692. https://doi.org/10.1016/j.procs.2024.04.253

[19] Singh, D. K., Kumar, A., Tamta, V. K., & Dutta, M. K. (2021). 3D-CNN based dynamic gesture recognition for Indian sign language modeling. Procedia Computer Science, 189, 76–83. https://doi.org/10.1016/j.procs.2021.05.071

[20] Gupta, R., & Rajan, S. (2020). Comparative analysis of convolution neural network models for continuous Indian sign language classification. Procedia Computer Science, 171, 1542–1550. https://doi.org/10.1016/j.procs.2020.04.165

[21] Suri, K., & Gupta, R. (2019). *Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory*. Computers & Electrical Engineering, 78, 493–503. https://doi.org/10.1016/j.compeleceng.2019.08.006

[22] Luciana, D., Alfena, C. A., Istianto, Y. R., & Anderies. (2025). Improving Indonesian sign language recognition using lightweight deep learning architectures with knowledge distillation method. *Procedia Computer Science, 269*, 618–629. https://doi.org/10.1016/j.procs.2025.09.005

[23] Rao, G. A., & Kishore, P. V. V. (2018). *Selfie video based continuous Indian sign language recognition system*. Ain Shams Engineering Journal, 9(4), 1929–1939. https://doi.org/10.1016/j.asej.2016.10.013

[24] Gupta, R., & Kumar, A. (2021). *Indian sign language recognition using wearable sensors and multi-label classification*. Computers & Electrical Engineering, 90, 106898. https://doi.org/10.1016/j.compeleceng.2020.106898.

[25] Tak, W., Arora, V., Chambyal, V., Kumari, V., Sumitra, & Chug, S. (2026). A dual-path deep learning framework for static and dynamic sign language recognition using CNN and LSTM architectures. *Procedia Computer Science, 283*, 2557–2566. https://doi.org/10.1016/j.procs.2026.06.325

[26] Mutlu, B., & Bilge, Y. C. (2026). An overview of sign language processing from natural language processing perspective. *Computer Science Review, 60*, 100907. https://doi.org/10.1016/j.cosrev.2026.100907

[27] Yerimbetova, A., Berzhanova, U., Sakenova, B., Milosz, M., Daiyrbayeva, E., Bayekeyeva, A., Mamyrbayeva, O., & Telman, D. (2026). Sign language recognition based on deep learning via MediaPipe for people with speech and hearing impairments. *Procedia Computer Science, 275*, 359–368. https://doi.org/10.1016/j.procs.2025.12.044

[28] Venkatesh, S., Kshirsagar, P. R., Thiagarajan, R., Tak, T. K., &Sivaneasan, B. (2026). Optimization model for sign language recognition using hybrid convolution networks. *Signal Processing: Image Communication, 141*, 117444. https://doi.org/10.1016/j.image.2025.117444

[29] Ismaiel, W., Kechiche, L., Aribi, Y., Omer, O. S. D., &Merghani, W. (2026). Leveraging edge detection techniques to enhance Arabic sign language static-gesture recognition using deep learning. *Journal of Engineering Research, 14*, 896–915. https://doi.org/10.1016/j.jer.2025.09.011

[30] Ashfaq, U., Wang, Q., Merabet, B., & Zhang, J. (2026). SignViT: An enhanced vision transformer framework for attention-based sign language hand gesture recognition. *Biomedical Signal Processing and Control, 112*, 108602. https://doi.org/10.1016/j.bspc.2025.108602

[31] Pranto, M. F. I., Islam, M. R., Akbor, M. A., Ghosh, N., Alam, M. R., Chaki, S., & Islam, M. M. (2026). A comprehensive image dataset of American Sign Language hand gestures. *Data in Brief, 65*, 112492. https://doi.org/10.1016/j.dib.2026.112492

[32] Ra, S. S., Sa, S., Ma, A. K. G., Siddhartha, P., &Nalinadevi, K. (2025). Sign language to sign language translator. *Procedia Computer Science, 260*, 373–381. https://doi.org/10.1016/j.procs.2025.03.213

[33] Tripathi, A., Kiran, D. S. V. N. R., Chandra, D. J., Kumar, M. B., Ram, A. S. D. S., Anjankar, S., Suthendran, K., Malviya N., Intelligent Sign Language Recognition for Real-Time Text Conversion to Aid Speech and Hearing Impaired, Procedia Computer Science, Volume 259, 2025, Pages 1472-1478, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2025.04.102.

[34] Kara, H., & Viswanathan, P. (2025). Nth layer hierarchical bidirectional LSTM sign language interpretation for hearing impaired person. *Procedia Computer Science, 258*, 3175–3183. https://doi.org/10.1016/j.procs.2025.04.575

[35] Ranee, L. A., Marshillong, F. L., Lyndoh, S. A., & Maji, A. K. (2025). Khasi sign language recognition using Google's MediaPipe and deep learning feedforward neural network approach. Procedia Computer Science, 258, 3619–3629. https://doi.org/10.1016/j.procs.2025.04.617

[36] Mishra, R., Angne, G., Gawde, N., Khamkar, P., &Utekar, S. (2025). SignSpeak: Indian Sign Language recognition with ML precision. *Indian Journal of Science and Technology, 18*(8), 620–634. https://doi.org/10.17485/IJST/v18i8.4049

[37] Sa, R., Rashmi, M., & Suresh, S. (2024). Sign language recognition by using spatio-temporal features. *Procedia Computer Science, 233*, 353–362. https://doi.org/10.1016/j.procs.2024.03.225

[38] Chaudhari, P., Vicente-Sola, A., Basu, A., Manna, D. L., Kirkland, P., & Di Caterina, G. (2024). Sign language recognition using spiking neural networks. *Procedia Computer Science, 235*, 2674–2683. https://doi.org/10.1016/j.procs.2024.04.252

[39] INCLUDE: A Large-Scale Dataset for Indian Sign Language Recognition https://dl.acm.org/doi/10.1145/3394171.3413528

[40] Latif, G., Alghazo, J., Alzubaidi, L., Naseer, M. M., &Alghazo, J. (2019). Arabic alphabets sign language dataset (ArASL) [Dataset]. Mendeley Data, V1. https://doi.org/10.17632/y7pckrw6z2.1

[41] Pranto, M. F. I., Islam, M. R., Akbor, M. A., Ghosh, N., Alam, M. R., Chaki, S., & Islam, M. M. (2026). ASL-HG: American Sign Language hand gesture image dataset [Dataset]. Mendeley Data, V1. https://doi.org/10.17632/j4y5w2c8w9.1

[42] Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A large-scale dataset and methods comparison. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 1459–1469). https://doi.org/10.1109/WACV45572.2020.9093360

[43] Joze, H. R. V., & Koller, O. (2019). MS-ASL: A large-scale data set and benchmark for understanding American Sign Language [Dataset]. Microsoft Research. https://doi.org/10.48550/arXiv.1812.01053

[44] Desai, A., Berger, M., Roche, M., &Pugeault, N. (2023). ASL Citizen: A community-sourced dataset for advancing isolated sign language recognition. arXiv. https://doi.org/10.48550/arXiv.2304.05934