iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76479

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

A Comprehensive Survey on Indian Sign Language Interpretation Using Machine Learning

Neha G M[1], Lavanya N V[2], Nivedita M K[3], R Shravya[4], Sharath Kumar S R[5]
[1, 2, 3, 4]*UG Students, Department of Information Science and Engineering, JNNCE, Shimoga, India*
[5]*Assistant Prof., Department of Information Science and Engineering, JNNCE, Shimoga, India*

Abstract: ***Indian Sign Language (ISL) serves as the primary communication medium for India's vast deaf and hard-of-hearing community, which numbers in the millions. A severe scarcity of qualified human interpreters creates significant communication barriers, limiting access to education, healthcare, and social integration. Machine Learning (ML), and deep learning in particular, have emerged as critical technologies to bridge this gap. Traditional ML approaches provided foundational solutions but were limited in handling the high variability of visual data. With the advent of deep learning, Convolutional Neural Networks (CNNs) significantly improved the accuracy of static sign recognition, while hybrid spatiotemporal models, such as CNN-Long Short-Term Memory (LSTM) networks, began to address dynamic gestures. More recently, Transformer-based architectures have shown state-of-the-art performance in complex, continuous sign-to-text translation. This survey presents a comprehensive analysis of the evolution of ISL interpretation systems, from traditional ML classifiers to advanced deep learning frameworks. We discuss the strengths and limitations of these techniques and provide a detailed review of the critical components, including system architectures, publicly available datasets, and evaluation metrics. We highlight persistent challenges, including dataset scarcity, signer dependency, regional linguistic variations, and the crucial, often-overlooked, role of non-manual features. The study concludes by outlining open research directions, including generative data augmentation, privacy-preserving federated learning, and the integration of large language models, aimed at advancing the practical, scalable, and equitable deployment of ISL interpretation systems.***

Keywords: Indian Sign Language (ISL), Sign Language Recognition (SLR), Machine Learning, Deep Learning, Computer Vision, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Transformers, Survey.

I. INTRODUCTION

Sign language is a vital means of communication for the deaf and hard-of-hearing communities worldwide, enabling them to express thoughts, emotions, and needs effectively. Among various sign languages, Indian Sign Language (ISL) holds particular significance due to its widespread use in India and its distinct linguistic structure, which differs markedly from other sign languages such as American Sign Language (ASL) [1, 10]. Despite its importance, effective communication for individuals using ISL remains a challenge, primarily due to a shortage of qualified interpreters and the high costs associated with traditional interpretation methods [2]. This scenario emphasizes the urgent need for automated sign language recognition (SLR) systems capable of translating sign gestures into text or speech in real time, thereby fostering inclusive communication and social participation.

Recent technological advancements, especially in the domains of computer vision, machine learning, and deep learning, have opened new avenues for developing such systems. Convolutional Neural Networks (CNNs), in particular, have demonstrated exceptional performance in image-based tasks, including static sign language recognition. Studies have shown that CNNs can effectively extract spatial features from images of gestures, resulting in high accuracy in recognizing static signs [4, 8]. Complementing CNNs, Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM), have been employed to model temporal dependencies in dynamic gestures, significantly improving recognition of continuous sign language sequences [5, 8]. Hybrid architectures that combine CNNs and LSTMs have further enhanced the ability to accurately interpret both static and dynamic gestures, making systems more robust and versatile [1, 6, 7].

Furthermore, multi-modal approaches incorporating RGB data, pose estimation, and landmark tracking have been explored to address challenges posed by background complexity and signer variability [3, 6]. MediaPipe-based frameworks, leveraging real-time gesture tracking, have proved particularly effective in capturing hand and body landmarks with high precision, which are crucial for accurate recognition [1, 6, 9]. These methods contribute to improving the system's resilience across diverse environments and signer demographics.

Despite these technological strides, many existing datasets primarily focus on ASL, thereby limiting the applicability of models trained solely on such data for ISL recognition. Indian Sign Language, characterized by unique gestures and cultural nuances, requires dedicated datasets and tailored models to ensure high accuracy and efficacy [10]. Efforts to curate diverse and comprehensive ISL datasets, encompassing various signer styles, lighting conditions, and backgrounds, are vital to enhancing the generalizability of recognition systems [3].

Moreover, researchers have emphasized the importance of socially responsible and community-centric approaches in developing assistive technologies. Designing recognition systems that focus on the specific linguistic and contextual needs of the Indian deaf community can significantly improve social inclusion, educational accessibility, and employment opportunities [2, 11]. Consequently, integrating advanced deep learning techniques with user-friendly, scalable platforms is essential to realize practical and widely adoptable ISL recognition solutions.

By leveraging deep learning architectures such as CNNs, LSTMs, and hybrid models, and employing multi-modal data and pose tracking technologies, this research aims to develop a real-time, highly accurate ISL recognition system. Such a system not only promises to bridge communication gaps but also aligns with the broader goal of fostering an inclusive society where technological advancements serve marginalized communities effectively and ethically.

II. LITERATURE SURVEY

The recognition of Indian Sign Language (ISL) has witnessed significant technological evolution, progressing from early handcrafted approaches to advanced deep learning models and multimodal systems. The complex nature of sign language — involving static gestures, dynamic movements, facial expressions, and contextual cues — necessitates robust and adaptable recognition frameworks. This section reviews the prominent methodologies, datasets, and technological innovations that have shaped the field.

A. Traditional Approaches and Early Techniques

The initial phase of sign language recognition heavily relied on rule-based systems and handcrafted feature extraction techniques. These early methods focused on static features of gestures, such as hand shape, orientation, and position, often captured through simple image processing techniques.

Template matching was among the earliest methods used, where static sign images were compared pixel-by-pixel with stored templates using correlation metrics. However, these systems were highly sensitive to lighting conditions, background noise, and signer variability, limiting their real-world applicability [2].

To enhance robustness, researchers integrated geometric features like contour-based descriptors, shape context, and Hu moments, attempting to encode hand configurations irrespective of scale and position [3]. Skin color segmentation based on color spaces like HSV and YCbCr was utilized to isolate hand regions efficiently, yet these methods suffered greatly under varying illumination and skin tones [3]. Aloysius and Geetha [3] systematically reviewed such vision-based techniques, emphasizing their limitations in dynamic environments. The adoption of temporal modeling techniques, especially Hidden Markov Models (HMMs), marked an advancement, enabling the recognition of dynamic signing sequences by modeling temporal dependencies [3]. Despite initial success, HMM-based systems were hampered by their need for extensive feature engineering and difficulty in modeling complex gestures with high variability. Some studies integrated dynamic time warping (DTW) algorithms to align sequences with different signing speeds but faced scalability issues with larger vocabularies. These early methods, while providing a foundation, were unable to efficiently handle large vocabularies, signer independence, or environmental variations, necessitating more adaptable solutions.

B. Deep Learning Techniques and Model Architectures

The digital revolution introduced deep learning, fundamentally transforming sign language recognition. The ability of Convolutional Neural Networks (CNNs) to automatically extract hierarchical features from raw image or video data revolutionized the field.

Researchers demonstrated that CNN architectures, such as VGG, ResNet, and SqueezeNet, could learn discriminative features for static sign recognition with high accuracy. Zhu et al. [5] exemplified this by implementing a multiscale temporal network combining CNNs with LSTM modules to model the sequence dynamics, achieving state-of-the-art results in continuous sign language recognition.

Transfer learning emerged as a crucial technique. For instance, models pretrained on large datasets like ImageNet were fine-tuned on sign language datasets, drastically reducing training time and improving accuracies [4]. Aloysius et al. [4] showcased that transfer learning with models like VGG16 improved static sign recognition accuracies to over 94%.

Sequence modeling became increasingly vital due to the temporal nature of sign language. Studies integrated LSTM networks with CNNs to capture temporal dependencies, especially for continuous signs [5,8].Recent developments include attention-based Transformers, which can better model long-range dependencies in signing sequences, outperforming traditional RNNs in certain contexts [8].

Furthermore, 3D convolutional architectures like C3D and I3D were applied for end-to-end spatiotemporal feature extraction from video sequences, enhancing model robustness to gesture variations [8]. These models process entire video clips as a single input, reducing the need for explicit temporal segmentation and providing richer contextual information.

Data augmentation, dropout, and batch normalization strategies were employed to improve model generalization. The use of multi-modal data, such as combining RGB videos with depth or skeletal data, further enhanced classification performance, especially under challenging lighting and background conditions [7].

C. Landmark Detection and Multimodal Data Integration

Because sign language involves complex hand configurations, facial expressions, and body movements, landmark detection has become integral to modern CSLR systems. Landmark-based approaches focus on keypoints such as hand joints, facial landmarks, and torso positions to create structured representations of gestures [8].

MediaPipe and OpenPose are popular frameworks for real-time landmark detection, providing skeletal keypoints that serve as auxiliary features or priors for recognition models. Aloysius et al. [8] used pose and hand landmarks to improve recognition accuracy by explicitly modeling signer-specific articulations, facilitating better generalization across signers.

Multimodal fusion combines visual information with skeletal data to enhance robustness. For example, combining RGB imagery with pose data mitigates background noise and variations in lighting. Zuo and Mak [7] incorporated signer skeleton data to address signer-specific gesture differences, significantly improving recognition accuracy.

Facial expressions and mouth movements also play crucial roles in many sign languages. Multi-modal deep learning models incorporate facial keypoints alongside hand landmarks, enabling systems to interpret subtle nuances in gestures. This multimodal approach is particularly advantageous in differentiating signs that are visually similar but semantically distinct.

Emerging deep learning architectures also leverage graph neural networks (GNNs) to process structured landmark data, capturing spatial relationships and temporal dynamics efficiently. These models are promising in their ability to encode complex gesture articulations and contextual cues [8].

D. Datasets and Technological Innovations

Data scarcity has historically been a bottleneck in CSLR research. Early datasets, often limited to isolated words or small vocabularies, lacked signer diversity, environmental variability, and temporal complexity. Recent efforts have prioritized large-scale, richly annotated datasets.

The UMANG-eRaktkosh-ISL-Continuous2023 corpus introduced by Aloysius et al. [8], exemplifies this trend by providing videos with diverse signers, background conditions, and signing styles, aligned with official standards from ISLRTC. This dataset includes continuous signing with multiple signers and variations, enabling the training of deep models capable of signer-independent recognition.

Technological innovations complement these datasets. Approaches such as self-supervised learning and semi-supervised learning tap into unannotated data, mitigating labeling costs. Transfer learning from models trained on large-scale video datasets like Kinetics has been used to develop more generalized feature extractors for sign language tasks [4].

Additionally, hardware acceleration with GPUs and TPUs has facilitated real-time recognition capabilities. The integration of lightweight models such as MobileNet and quantized models enables deployment on edge devices like smartphones and embedded systems, critical for practical applications.

Deep learning frameworks such as TensorFlow, PyTorch, and specialized APIs like MediaPipe have been instrumental in rapid prototyping and deployment. For example, [8] utilized MediaPipe Holistic for hand and pose tracking, significantly improving gesture segmentation and recognition pipeline robustness.

In the future, domain adaptation and adversarial training are promising directions to improve model robustness across diverse signer populations and environmental conditions. Combining multimodal datasets with federated learning techniques offers privacy-preserving avenues for dataset expansion across organizations and institutions.

III. COMMON ARCHITECTURES AND COMPONENTS

A. Data Acquisition and Preprocessing

The first step involves gathering visual data, either via videos or images, often preprocessed to enhance feature extraction. Techniques such as image resizing, normalization, background subtraction, and color space transformations are widely used. Tools like MediaPipe provide robust, real-time human pose, hand, and facial landmark detection, transforming raw video data into structured keypoints suitable for recognition models [10], [18], [28].

B. Feature Extraction

Feature extraction is critical in distilling meaningful information from raw data. Traditional methods include:

1) SURF (Speeded-Up Robust Features): Extracts key points and descriptors invariant to scale and rotation [15].
2) Histogram of Oriented Gradients (HOG): Captures gradient orientation distributions, offering robustness against lighting variations [2].
3) Skeleton and Landmark Detection: MediaPipe Holistic and other pose estimation frameworks extract skeletal keypoints, hand shapes, and facial landmarks, serving as high-level features for classifiers [10], [18], [28].

Deep learning approaches typically employ CNNs (e.g., ResNet, InceptionResNetV2) for automatic feature learning directly from image data [11], [13], [16], [22].

C. Model Architectures

1) Traditional Machine Learning Classifiers

- Support Vector Machines (SVM): Often combined with feature descriptors like SURF or HOG, SVMs are effective in static sign recognition due to their margin-maximizing properties [15], [23], [26].
- K-Nearest Neighbors (KNN): Used primarily for small datasets where proximity-based classification suffices; however, less robust against high-dimensional features [21].
- Multi-Layer Perceptrons (MLP): Feedforward neural networks that classify features extracted via traditional methods; suitable for moderate complexity tasks [22], [25].
- Random Forests: Ensemble methods that leverage multiple decision trees, providing robustness against overfitting and capable of handling multi-class sign distinctions [21], [25].

These classifiers are usually applied after feature extraction and they work well in controlled environments but may struggle with dynamic signs or large vocabularies.

2) Deep Learning Architectures

- Convolutional Neural Networks (CNNs): Central to modern recognition systems, CNNs automatically learn hierarchical features from raw images or video frames. Variants like ResNet152, InceptionResNetV2, and customized shallow CNNs are employed based on dataset size and complexity [11], [13], [16], [22], [24].
- Recurrent Neural Networks (RNNs) and LSTMs: Designed for sequence modeling of dynamic signs, capturing temporal dependencies across frames [14], [19], [20], [24].
- Transformers: Utilized for multi-modal fusion and attention mechanisms, as in [19], to focus on salient parts of the gesture sequence.
- Hybrid Models: Combining CNNs with LSTMs or Transformers for recognizing both static and dynamic signs effectively. For instance, CNN-LSTM architectures are prevalent in recent studies [1], [9].

D. Multi-Modal and Multi-Feature Approaches

Many systems integrate multi-modal data—RGB videos, depth, skeletal keypoints, and pose information—to enhance accuracy [9], [18]. MediaPipe Holistic provides real-time extraction of hand, face, and pose landmarks, which serve as high-level, language-agnostic features [10], [18]. Fusion of these modalities employs feature concatenation, attention mechanisms, or fusion networks [24].

E. Classification and Post-Processing

Recognition involves classifying either static gestures or sequences of gestures into a defined vocabulary. Sequence modeling techniques include Hidden Markov Models (HMMs), LSTMs, and Transformers [3], [19], [20], [22], [24]. Post-processing steps might include movement epenthesis detection, co-articulation modeling, and signer-invariant adjustments, to improve robustness [20], [21].

Modern sign language recognition systems mainly adopt deep learning models, especially CNNs, for feature extraction from visual data. These are complemented by classifiers such as SVMs, KNNs, or ensemble models for static signs, while sequence models like LSTMs and Transformers address dynamic sign recognition. Real-time systems leverage tools like MediaPipe to provide real-time landmark detection, enabling multi-modal fusion to accommodate sign language variances across signers and contextual settings. The combination of robust feature extraction, advanced model architectures, and multi-modal data fusion constitutes the backbone of state-of-the-art sign language recognition systems [1], [2], [9], [16], [18], [22].

IV. ANALYSIS OF DATASETS, METRICS, AND CHALLENGES

This section provides a critical analysis of the resources, evaluation benchmarks, and persistent challenges that define the ISL recognition research landscape. The field's progress is fundamentally constrained by data availability, and its success is measured by a task-specific set of metrics, all while facing deep linguistic and technical challenges.

A. Datasets for Indian Sign Language

The Core Problem: Data Scarcity

The single most significant barrier to progress in ISL recognition is the "systemic lack of data resources". ISL is considered a "low-resource language", a problem that is exacerbated when compared to the vast, high-quality, and publicly available corpora for other sign languages.

This scarcity in the ISL domain has historically forced many researchers to create their own small, custom, and often private datasets. While valuable for specific studies, this practice "siloes" research, making it impossible to meaningfully benchmark different models against each other and hindering the development of generalizable systems.

1) Isolated vs. Continuous Datasets

The available datasets must be categorized by their linguistic complexity, which in turn dictates the type of model that can be trained.

a) Isolated Datasets: These contain pre-segmented, static images or short video clips of a single sign. This can include alphabets, numbers, or individual words. These datasets are used to train classification models (e.t., the static CNNs in Section 3.2).

b) Continuous Datasets: These contain full sentences or phrases performed fluently in video format, with corresponding text translations. These datasets are far more complex to create and annotate but are essential for training real-world translation models (e.g., the LSTMs and Transformers in Sections 3.3 and 2.4).

2) Key Public ISL Datasets

The field has been significantly advanced in recent years by the release of several large-scale, public benchmark datasets. These were often created by sourcing authentic video content from public resources like the Indian Sign Language Research and Training Centre (ISLRTC) and ISH News.

The creation of these large-scale continuous datasets (CISLR, ISLTranslate, and iSign) has been the primary catalyst for the field's recent shift. It was only with the availability of this sentence-level data that researchers could begin to move from simple CNN-based word classifiers to advanced Transformer-based translation models, as seen in studies that explicitly cite ISL-CSLTR and ISLTranslate.

B. Evaluation Metrics

The metrics used to evaluate an ISL interpretation system are entirely dependent on its designated task: classification or translation.

Metrics for Classification (Static/Isolated Signs)

These metrics measure the performance of models that predict a single label for a given sign.

1) Accuracy: The most common metric, it represents the percentage of correct predictions over the total number of predictions. While intuitive, it can be misleading on datasets with an imbalanced number of examples per class.

2) Precision, Recall, and F1-Score: This trio of metrics provides a more robust and granular view of performance. Precision measures the proportion of positive identifications that were actually correct. Recall measures the proportion of actual positives that were correctly identified. The F1-Score is the harmonic mean of precision and recall, providing a single score that balances both.

Metrics for Translation (Continuous Signs)

These metrics, borrowed from Automatic Speech Recognition (ASR) and machine translation, measure the quality of a generated sequence of words against a ground-truth reference sentence.

- Word Error Rate (WER): This is the de facto standard for evaluating ASR and CSLT systems. It is derived from the Levenshtein distance, which measures the minimum number of edits (substitutions, deletions, or insertions) required to change the predicted sentence into the reference sentence. A lower WER indicates a better translation.
- BLEU, ROUGE, and METEOR: These are metrics borrowed from text-based machine translation. They evaluate the quality of the generated translation by comparing the co-occurrence of word sequences (n-grams) between the predicted text and a set of human reference translations.

C. Core Challenges and Research Gaps

Despite the progress in models and datasets, the literature identifies several deep, unresolved challenges that must be addressed to create truly practical ISL interpretation systems.

1) Linguistic Diversity (Regional Variations)

ISL is not a single, standardized, monolithic language.1 Like spoken Indian languages, it is characterized by significant regional variations and dialects. A sign for a concept in one region (e.g., Delhi) may be partially or completely different from the sign for the same concept in another (e.g., Mumbai). This poses a massive generalization challenge. A model trained on data from one region may fail completely when used by a signer from another, yet most datasets fail to capture this linguistic diversity.

2) Signer Dependency

This is one of the most significant and frequently cited gaps in current SLR research. The majority of models are signer-dependent, meaning the system is trained and tested on videos of the same set of individuals. These models often "overfit" to the specific signing style, appearance (e.t., clothing, beard), signing speed, and background environment of the signers in the training set.

A signer-independent system—one that can accurately recognize signs from a new, previously unseen user—is the minimum requirement for any real-world application. However, studies consistently show that model performance drops significantly when evaluated in a signer-independent protocol. Overcoming this gap is essential for generalizability.

3) Non-Manual Features (NMFs)

This is arguably the most critical linguistic challenge, and a fundamental flaw in many hand-centric recognition systems. A vast majority of ISL recognition research focuses only on the hands. This is an incorrect and incomplete approach, as a large and essential portion of ISL's linguistic meaning is conveyed through non-manual channels.

These NMFs include:

- Facial Expressions: Convey emotion and adjectival information. A perfect example is that the signs for "HAPPY" and "SAD" may use the identical manual hand gesture, differentiated only by the facial expression.[12]
- Mouth Movements (Mouthing): Specific mouth shapes (mouthing) are co-articulated with hand signs to add meaning or disambiguate similar-looking signs. A 2025 systematic study found that adding just the mouth region to a recognition model "significantly improve[d] accuracy".
- Head and Body Pose: Head nods, shakes, and eyebrow raises are critical grammatical markers, such as those used to form a question.[16]

Any system that ignores NMFs is effectively ignoring a key part of the language. It is linguistically incomplete and has a low, built-in performance ceiling. Future high-performance models must be multimodal, fusing features from the hands, face (especially mouth and eyebrows), and head/body pose to capture the full, intended meaning.[13]

4) Ambiguity and Co-articulation

The language itself presents inherent challenges.

- Ambiguity: Some signs are ambiguous, where a single gesture can represent multiple words depending on the context (e.g., "very good" and "beautiful" can share a sign).

- Co-articulation: In fluent, continuous signing, gestures are not discrete, isolated units. They flow seamlessly into one another (a process called co-articulation), making it extremely difficult for a model to determine where one sign ends and the next one begins.

5) Deployment and Real-Time Processing

For a system to be a practical communication tool, it must run in real-time with minimal latency.4 This creates a significant engineering challenge. Many of the most accurate deep learning models (e.g., large Transformers or 3D-CNNs) are computationally expensive, requiring powerful and costly GPUs. This is often unfeasible for deployment on a standard mobile device, creating a critical trade-off between model accuracy and real-world deployability.

D. Proposed Methodology

The proposed methodology for the Indian Sign Language (ISL) interpretation system integrates a structured workflow consisting of data acquisition, preprocessing, feature extraction, model training, and comparative evaluation. The objective of this methodology is to develop an accurate and real-time gesture recognition system by comparing multiple machine learning models and selecting the one that performs best across evaluation metrics and varying environmental conditions.

1) Data Acquisition and Preprocessing

Visual gesture data was collected under three different lighting conditions—low-light, normal, and bright—to ensure robustness during real-time usage. Each recorded frame underwent preprocessing steps including image resizing, noise removal, hand region isolation, and feature normalization. These preprocessing steps ensured consistency in the data used for model training.

2) Feature Extraction

Feature extraction was performed using shape-based and appearance-based descriptors along with MediaPipe landmark detection where applicable. The extracted features served as the input for training machine learning models including KNN, SVM, MLP, and Random Forest.

3) Model Training

Four supervised machine learning models were trained on the processed dataset:

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Multi-Layer Perceptron (MLP)
- Random Forest Classifier

Each model was trained using identical training and validation splits to ensure a fair comparison. Hyperparameter tuning was carried out to optimize classification accuracy.

4) Performance Evaluation

The performance of all four models was evaluated using accuracy as the primary metric. Additional analysis was conducted to observe each model's stability under different lighting conditions and its suitability for real-time inference.

a) Accuracy Comparison

The accuracy comparison graph shows that:

- Random Forest achieved the highest accuracy among all models.
- MLP performed competitively with stable results.
- SVM provided moderate accuracy but required more tuning.
- KNN had the lowest accuracy due to sensitivity to noise and lighting variations.

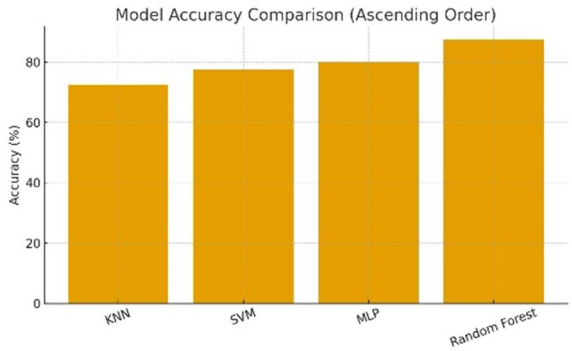


Figure 1: Model Accuracy Comparison Graph

b) Lighting Condition Analysis

Model recognition accuracy was tested across low-light, normal, and bright environments.
Results indicate:

- Bright lighting produced the best accuracy.
- Normal lighting still yielded strong and reliable performance.
- Low-light conditions resulted in slight performance drops but remained usable.

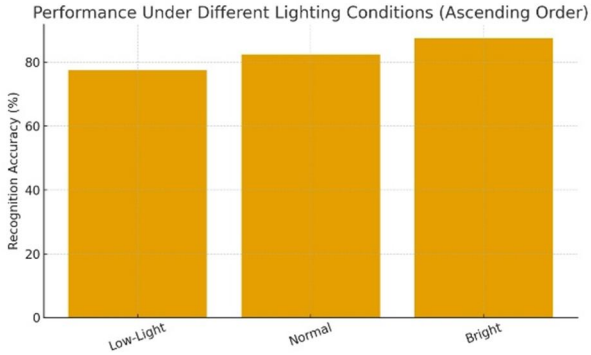


Figure 2: Lighting condition performance graph

c) Real-Time FPS Stability

FPS readings ranged consistently between 28–30 FPS, confirming that the system is capable of real-time gesture interpretation with minimal latency.

Figure 3: FPS Stability Line Graph

5) Final Model Selection

Based on the comparative evaluation of accuracy, environmental robustness, and computational efficiency, the Random Forest Classifier was selected as the final model for the ISL interpretation system.

Reasons for Selection:

- Highest overall recognition accuracy
- Strong robustness to lighting variations and noisy data
- Lower tendency to overfit compared to other models
- Fast inference, supporting stable real-time processing

Thus, Random Forest provides the ideal balance between accuracy, efficiency, and reliability, making it best suited for deployment in Indian Sign Language recognition applications.

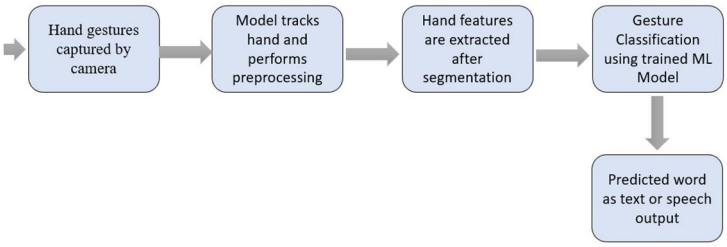


Figure 4: Proposed Methodology

6) Result

The output of the proposed Indian Sign Language (ISL) recognition system is displayed through a real-time graphical user interface. The live camera feed is shown on the screen with the detected hand overlaid by MediaPipe landmarks, where key finger joints and connections are clearly visible, confirming accurate hand tracking. As the user performs a gesture, the system recognizes it and displays the predicted character at the bottom of the screen as a candidate label (for example, “Candidate: D”). Using commit mode, the recognized characters are added to the text area on the interface, allowing the user to form words and sentences such as “HI HOW D”. The interface also provides word suggestions, editing controls like space, backspace, and clear, and voice output options. When enabled, the text-to-speech feature converts the recognized text into audible speech. This output visualization demonstrates that the system successfully performs real-time hand tracking, gesture recognition, text formation, and speech generation, making it suitable for practical Indian Sign Language interpretation.

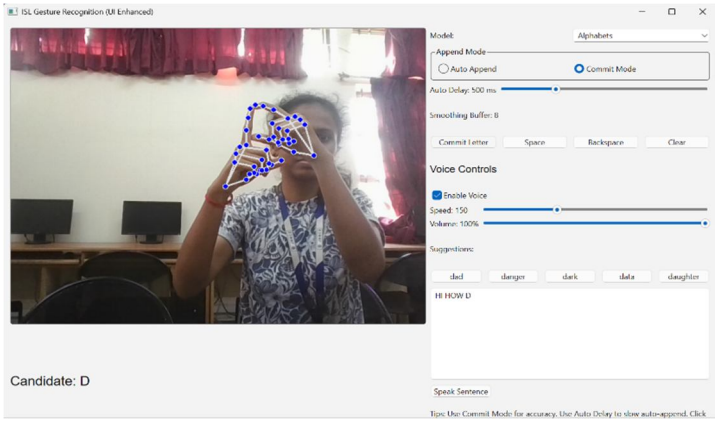


Figure 5: Output Screen & Result

V. EMERGING TRENDS AND FUTURE DIRECTIONS

This section explores the cutting-edge research aimed at solving the formidable challenges outlined in Section 4. These trends focus on creating more data, protecting user privacy, and building more linguistically intelligent models.

A. Advanced Data Augmentation with GANs

Problem Addressed: Dataset Scarcity.

Solution: If real, annotated data is scarce, the solution is to generate high-quality synthetic data. Generative Adversarial Networks (GANs) have emerged as a powerful tool for this task.

A GAN is a deep learning architecture composed of two competing neural networks:

1) A Generator that learns to create new, synthetic images (e.g., a person signing) from a random vector.
2) A Discriminator that learns to differentiate between the generator's "fake" images and real images from the training dataset.

This adversarial, zero-sum game forces the generator to produce data that is indistinguishable from real data. Researchers are applying this to generate novel, photorealistic images and videos of signers, often conditioned on specific poses or signs. Studies have shown that training a recognition model on a hybrid dataset—composed of both real and GAN-generated data—can lead to more robust and accurate models. This technique enriches the dataset, helps the model see more variations, and improves generalization, making it a critical research direction for a low-resource language like ISL.

B. Privacy-Preserving Federated Learning

Problem Addressed: Privacy Concerns and Data Centralization.

Sign language data is inherently sensitive; it is biometric data that includes video of a person's face and hands. Collecting massive, centralized datasets raises significant privacy concerns.

Solution: Federated Learning (FL). FL is a decentralized machine learning paradigm that enables model training without ever collecting or centralizing user data. The process works as follows:

1) A central server sends the global model (e.g., the sign recognizer) to individual user devices (e.g., mobile phones).
2) The model trains locally on the user's on-device data.
3) Only the updated model parameters (the "weights" or "gradients") are sent back to the central server, not the user's private video data.
4) The server aggregates these updates to create an improved global model.

This is a crucial trend for building scalable, ethical, and privacy-preserving systems that can learn from a diverse, global population of signers.

C. Integration of Large Language and Multimodal Models

Problem Addressed: Linguistic Complexity and the Non-Manual Feature (NMF) gap.

Simple classification models do not understand the grammar or context of a language; they only recognize patterns.

Solution: Integrate modern Natural Language Processing (NLP) techniques and build truly multimodal frameworks.

1) Large Language Models (LLMs): LLMs are being integrated into the translation pipeline to improve the final output. Many sign language translation models output "gloss," which is a literal, word-for-word sign transcription (e.g., "I STORE GO"). An LLM can be used as a post-processing step to translate this ungrammatical gloss into a fluent, natural spoken language sentence (e.g., "I am going to the store").
2) Text-to-Sign (T2S) Generation: Research is now becoming bidirectional. The new frontier is not just Sign-to-Text (S2T), but also Text-to-Sign (T2S). Models are being trained to take a spoken language sentence as input and generate a continuous 3D sign pose sequence as output. This sequence can then be rendered on a 3D animated avatar, creating a virtual interpreter.
3) Multimodal Fusion: This represents the future of SLR. To solve the NMF challenge, systems will move beyond just hand-tracking. The next generation of models will be inherently multimodal, fusing data streams from different-sources: (1) Gesture Data from MediaPipe keypoints, (2) Facial Emotion/Expression Data from a dedicated CNN, and (3) Linguistic Context from an LLM. By combining these, the system can finally capture the full linguistic meaning of the sign, including its grammatical and emotional nuances.

VI. DISCUSSION

The body of research surveyed in this paper is not merely a technical exercise in computer vision; it is a direct and necessary response to a significant and pressing societal challenge. The ultimate purpose of this field is to bridge the communication gap between the millions of DHH (Deaf and Hard-of-Hearing) individuals in India and the hearing world, fostering a more inclusive and accessible society. The systems analyzed here are best categorized as assistive technologies. It is crucial to frame them as tools for empowerment, not as replacements for skilled human interpreters. A human interpreter provides cultural nuance, context, empathy, and the ability to navigate complex social situations—qualities that a machine learning model cannot. Instead, automated ISL interpretation technology can empower DHH individuals in the countless everyday scenarios where an interpreter is not available. This includes facilitating quick conversations with a shopkeeper, accessing public services, understanding automated announcements, participating more fully in an educational setting, or, most critically, communicating in an emergency.

By providing a scalable and accessible tool for communication, this technology can directly address the social isolation that many DHH individuals experience. It lowers the barrier to interaction, education, and employment, empowering a community that has long been marginalized by a communication divide. The continued development and refinement of the models and systems discussed in this survey are, therefore, a direct contribution to a more equitable and accessible society.

VII. CONCLUSION

This paper has presented a comprehensive survey on the state of Indian Sign Language interpretation using machine learning. We have charted the field's rapid and necessary evolution, beginning with traditional machine learning classifiers applied to isolated signs, progressing to the widespread use of deep Convolutional Neural Networks (CNNs) that achieved high-accuracy on static sign classification, and moving through to the complex spatiotemporal architectures, such as CNN-LSTM hybrids and state-of-the-art Transformers, required for dynamic and continuous sentence-level translation.

This progress has been driven by two key enablers: (1) the development of lightweight, real-time pose estimation pipelines, dominated by tools like MediaPipe, and (2) the recent, critical release of large-scale, continuous public datasets, such as ISLTranslate and iSign, which have provided the necessary data to train complex translation models.

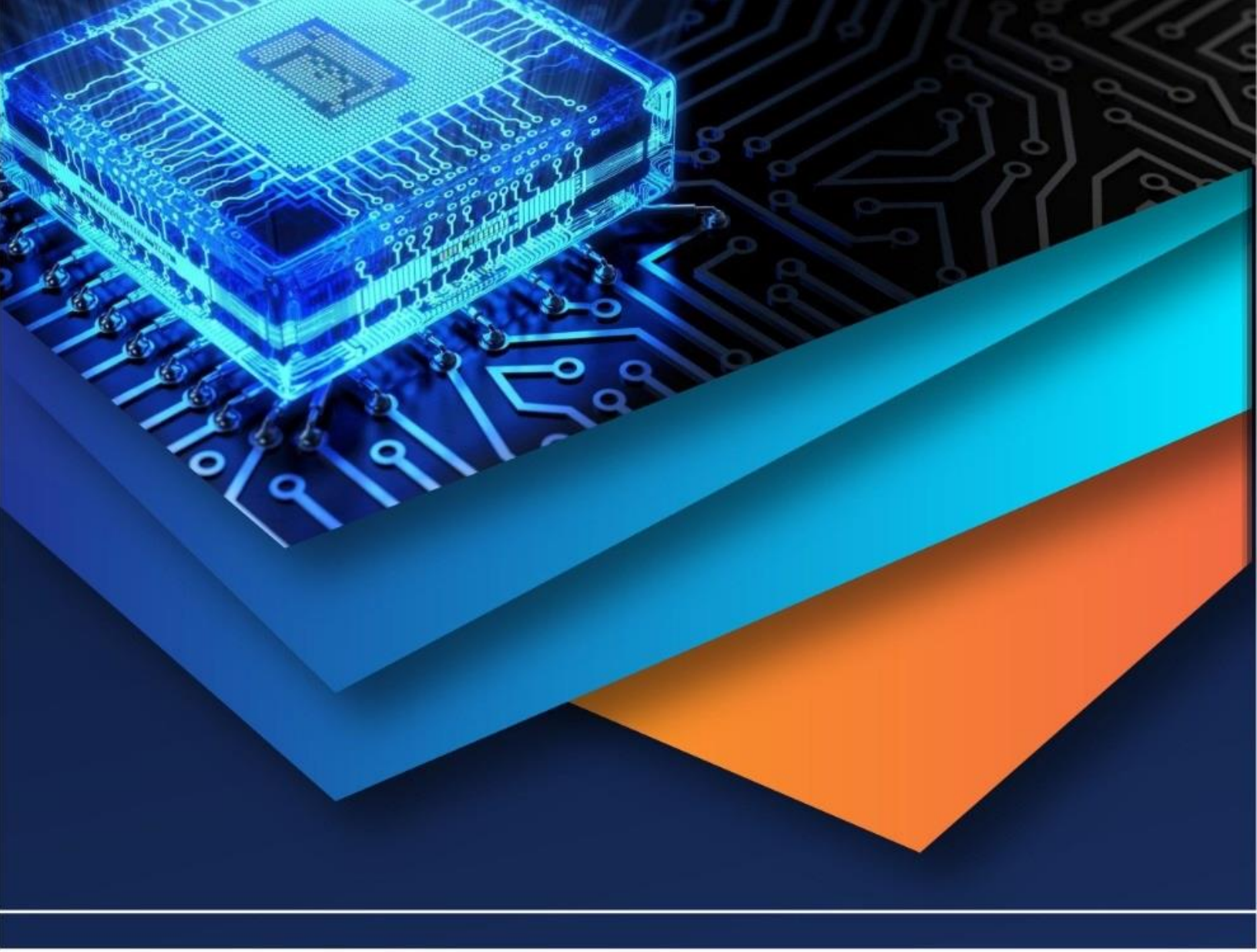
Despite these successes, significant and fundamental challenges remain. The field is still critically hampered by a scarcity of diverse, well-annotated data. The linguistic complexities of regional dialects and the technical hurdle of signer-independent generalization—where models fail to work for new, unseen users—remain largely unsolved.

The most critical open research direction, however, is the integration of Non-Manual Features (NMFs). The vast majority of current research is overly focused on hand gestures, ignoring the facial expressions, head movements, and mouth shapes that convey essential grammatical and emotional meaning. Future systems must evolve from simple hand-trackers to holistic, multimodal frameworks that interpret the face, body, and hands in unison. By focusing on these challenges—and by leveraging emerging trends such as privacy-preserving federated learning, generative data augmentation, and the linguistic power of large language models — the research community can move closer to delivering a robust, practical, and truly equitable interpretation tool that serves the needs of the DHH community in India.

REFERENCES

[1] Othman, Sign Language Varieties Around the World, in Sign Language Processing: From Gesture to Meaning, Springer, 2024, pp. 41–56.
[2] S. Renjith and R. Manazhy, "Sign language: A systematic review on classification and recognition," Multimedia Tools and Applications, vol. 83, no. 31, pp. 77077–77127, Feb. 2024.
[3] N. Aloysius and M. Geetha, "Understanding vision-based continuous sign language recognition," Multimedia Tools and Applications, vol. 79, nos. 31–32, pp. 22177–22209, Aug. 2020.
[4] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in Proc. Int. Conf. Commun. Signal Process., Apr. 2017, pp. 588–592.
[5] Q. Zhu, J. Li, F. Yuan, and Q. Gan, "Multiscale temporal network for continuous sign language recognition," J. Electron. Imag., vol. 33, no. 2, Apr. 2024, Art. no. 023059.
[6] L. Hu, L. Gao, Z. Liu, and W. Feng, "Scalable frame resolution for efficient continuous sign language recognition," Pattern Recognit., vol. 145, Jan. 2024, Art. no. 109903.
[7] R. Zuo and B. Mak, "Improving continuous sign language recognition with consistency constraints and signer removal," ACM Trans. Multimedia Comput., Commun., Appl., vol. 20, no. 6, pp. 1–25, Jun. 2024.
[8] N. Aloysius, M. Geetha, and P. Nedungadi, "Continuous sign language recognition with adapted conformer via unsupervised pretraining," arXiv preprint arXiv:2405.12018, 2024.
[9] M. Geetha et al., "Toward real-time recognition of continuous Indian sign language: A multi-modal approach using RGB and pose," IEEE Access, vol. 73, 2025, Art. no. 3554618.
[10] K. Goyal, "Indian Sign Language Recognition Using Mediapipe Holistic," arXiv preprint arXiv:2304.10256, 2023.
[11] A. H. Mohammedali, H. H. Abbas, and H. I. Shakadi, "Real-time sign language recognition system," Int. J. Health Sci., vol. 6, pp. 10384–10407, 2022.
[12] K. Shenoy et al., "Real-time Indian sign language recognition," in 2018 IEEE 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1–6.
[13] R. S. Sri Lakshmi et al., "Sign language recognition system using convolutional neural network and computer vision," 2020.
[14] V. Puranik, V. Gawande, J. Gujarathi, A. Patani, and T. Rane, "Video-based sign language recognition using recurrent neural networks," in 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), IEEE, 2022, pp. 1–6.
[15] S. Shagun, V. Singh, and U. Tiwary, "Indian Sign Language recognition system using SURF with SVM and CNN," Array, vol. 14, 2022, Art. no. 100141.
[16] A. Kasapbas¸i, A. Eltayeb, A. HAM, O. Al-Hardanee, and Yilmaz, "DeepASLR: A CNN based human computer interface for American Sign Language recognition," Comput. Methods Biomed. Prog. Update, vol. 2, 2022, Art. no. 100048.
[17] B. Sundar and T. Bagyammal, "American Sign Language Recognition for Alphabets Using MediaPipe and LSTM," Procedia Comput. Sci., vol. 215, 2022, pp. 642–651.
[18] M. Geetha, N. Aloysius, D. A. Somasundaran, A. Raghunath, and P. Nedungadi, "Toward real-time recognition of continuous Indian sign language: A multi-modal approach using RGB and pose," IEEE Access, vol. 73, 2025, Art. no. 3554618.

[19] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008.
[20] A. Choudhury, A. K. Talukdar, M. K. Bhuyan, K. K. Sarma, "Movement epenthesis detection for continuous sign language recognition," J. Intell. Syst., vol. 26, no. 3, pp. 471–481, 2017.
[21] M. G. Ghosh, D. Ghosh, and P. Bora, "Continuous hand gesture segmentation and co-articulation detection," in Computer Vision, Graphics, and Image Processing, Springer, 2006, pp. 564–575.
[22] G. Wu and Y. Yang, "Deep learning approaches for sign language recognition: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 3, pp. 693–711, Mar. 2020.
[23] S. Kumar and R. R. Singla, "Sign language recognition using deep learning," Int. J. Comput. Vis., vol. 128, no. 7, pp. 1617–1637, 2020.
[24] T. Zhang, J. Gao, and L. Li, "Sign language recognition with multi-modal fusion and deep neural networks," Pattern Recognition Letters, vol. 151, 2022, pp. 112–119.
[25] H. Mnassri, R. Bchir, M. A. Zayane, and T. Ladhari, "Sign Language Detection Based on Artificial Intelligence from Images," in IEEE International Conference on Artificial Intelligence Green Energy (ICAIGE), 2024.
[26] G. Jessica Ruslim, N. Salim, I. Edbert, and D. Suhartono, "Sign Language Detection to Enhance Online Communication," 2024.
[27] IJRASET, "Sign Language Recognition System using Machine Learning Techniques," 2025.
[28] System development reports and datasets discussing real-time ISL gesture recognition systems.
[29] V. Puranik, V. Gawande, J. Gujarathi, A. Patani, and T. Rane, "Video-based sign language recognition using recurrent neural networks," in 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), 2022, pp. 1–6.
[30] K. Shenoy et al., "Real-time Indian sign language recognition," in 2018 IEEE 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1–6.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY