Kateb
Scientific–Research Journal of Technology and Engineering
Vol.1 | No.1 | Autumn & Winter 2025–2026
Kateb University
Email: kjte@kateb.edu.af | **Homepage:** kjte.kateb.edu.af

Development of Gesture Recognition System for Converting Indian Sign Language Videos to Speech

Muhammad Wares Muhammadi[1]✉ **Reshad Ahmad Hussaini**[2]

1. Faculty Member, Faculty of Computer Science, Kateb University, Kabul, Afghanistan. (Corresponding Author) E-mail: Ahmadwares2016@gmail.com
2. Visiting Lecturer, Faculty of Computer Science, Kateb University, Kabul, Afghanistan. Email: reshada022@gmail.com

Article Info

Article type:
Research Article

Article history:
Received: 02/02/2026
Received in revised form: 21/02/2026
Accepted: 12/03/2026
Available online: 19/03/2026

Keywords:
Indian sign language (ISL), Speech conversion, Self-trained model, Deaf and hard-of-hearing community, Accessibility

ABSTRACT

In a world marked by diverse communication methods, the gap between hearing and deaf communities remains a significant challenge. Millions who rely on Indian Sign Language (ISL) face barriers in education, employment, and social interaction. This paper presents a machine learning–based gesture recognition system designed to convert ISL videos into text and speech in real time. Unlike systems that rely solely on pre-trained models, the proposed approach utilizes a self-trained and dynamically adaptive model, built using annotated ISL video data and continuously improved through user interaction. The system addresses key challenges such as dataset collection, ethical considerations, and the complexity of interpreting hand gestures, facial expressions, and body movements. By integrating computer vision techniques with deep learning models and text-to-speech synthesis, the system achieves high accuracy and real-time performance. This work contributes toward developing an inclusive communication framework that enhances accessibility and bridges the communication gap for individuals relying on ISL.

Cite this article: Muhammadi, M. & Hussaini, R. (2026). Development of Gesture Recognition System for Converting Indian Sign Language Videos to Speech, *Kateb Scientific-Research Journal of Technology and Engineering, 1 (1),* 1-22.

 Publisher: Kateb University.

Introduction

Envisioning a vibrant community where communication effortlessly bridges the gap between spoken and signed languages remains an elusive reality for millions who rely on Indian Sign Language (ISL) [3]. The profound disparity between the hearing and deaf and hard-of-hearing communities presents significant obstacles across various domains, including education, employment, and social interactions.

The emergence of automatic conversion systems holds promise as a means to narrow this gap. By harnessing the capabilities of machine learning, these systems seek to translate the intricate gestures of ISL into spoken language, facilitating seamless communication and enhancing accessibility. A particularly promising approach involves the utilization of self-trained models, which can learn and adapt from extensive datasets of ISL videos.

However, the development of such a system presents numerous challenges. Ethical considerations loom large in the acquisition of a diverse and comprehensive dataset, while accurately deciphering the nuanced combination of hand gestures, facial expressions, and body language inherent in ISL requires sophisticated algorithms. Furthermore, achieving real-time processing introduces computational hurdles that demand innovative solutions.

Figure 1 illustrates a set of standard Indian Sign Language (ISL) hand poses representing digits (0–9) and alphabetic gestures (A–Z). These hand configurations form the foundational visual vocabulary used in ISL and are essential for training gesture recognition models. The diversity and subtle variations among these poses highlight the complexity involved in accurately detecting and classifying gestures in real-time systems.

This paper delves into the potential of a self-trained model for real-time ISL to speech conversion in India. We thoroughly examine the challenges, technical intricacies, and potential impact of such a system on the lives of millions [4]. By addressing these challenges and leveraging the power of machine learning, our goal is to contribute to a world where communication transcends barriers.

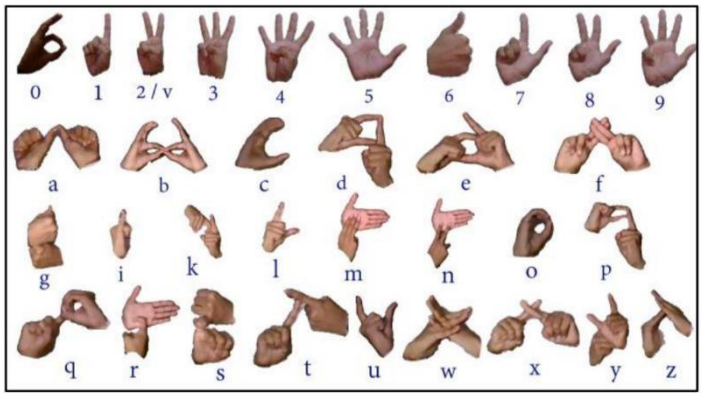

Fig. 1. Hand poses in ISL

This paper presents a system that accurately classifies the 33 single-handed hand poses within Indian Sign Language (ISL). While the initial focus was on one-handed

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

gestures, the system's design allows for effortless expansion to include two-handed gestures.

Related Work

The quest to bridge the communication gap between deaf and hearing communities has spurred a vibrant field of research in sign language translation. This endeavor encompasses diverse approaches applicable to various sign languages, including Taiwanese Sign Language (TSL), American Sign Language (ASL), Mexican Sign Language (Lengua de Señas Mexicana - LSM), Arabic Sign Language, Spanish Sign Language (Lengua de Signos Española - LSE), and Indian Sign Language (ISL). [5]

Delving into the Techniques:

Researchers have devised a rich tapestry of techniques to tackle sign language translation:

- ***Sensor-Embedded Gloves:*** These innovative gloves capture intricate hand movements, translating them into spoken language or text, offering a direct and potentially intuitive communication method. [6]
- ***Mobile Accessibility:*** Mobile applications leverage image processing and extensive sign language databases to translate signs recognized through the device's camera. [7] This approach provides a convenient and portable solution for everyday communication.
- ***Machine Learning for Empowerment:*** Machine learning algorithms have revolutionized the field by enabling automated translation from signs to text. This advancement significantly enhances accessibility for deaf individuals, fostering greater independence and participation in various domains. [8]
- ***Bridging the Spoken and Signed Divide:*** Research efforts have explored speech-to-sign and sign-to-speech translation systems. These systems utilize a combination of techniques like automatic speech recognition, animation, statistical modules, and word-to-speech conversion, facilitating seamless communication between deaf and hearing individuals.
- ***Motion Capture for Fidelity in Sign Language Depiction:*** Motion capture technology has found its place in sign language translation by faithfully capturing and reproducing sign language movements. [9] This technology holds promise for communication assistance and language learning applications.

The aforementioned techniques represent a snapshot of the ongoing efforts to bridge the communication gap between deaf and hearing communities. [10] This research paves the way for further advancements in sign language translation systems, particularly focusing on ISL, to enhance accessibility and inclusivity for millions in India.

Implementation

In the development of our gesture recognition system for converting Indian Sign Language (ISL) videos to speech, meticulous attention is paid to data cleaning and

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

pre-processing. Before training the dynamic ISL model, raw video data undergoes rigorous cleaning to remove noise and irrelevant information, ensuring the quality and consistency of the dataset. Subsequently, pre-processing steps such as resizing, normalization, and augmentation are applied to standardize the input data and enhance model robustness. [11] For user input, our system seamlessly accepts video frames and audio recordings, empowering users to upload ISL videos featuring sign language gestures alongside corresponding audio recordings of spoken words or phrases. This rich input data serves as the training dataset for our dynamic ISL model and acts as the source for generating scripts from audio using Google Speech Recognition. [12] Training the dynamic ISL model occurs in real-time as users interact with the system, enabling continuous learning and adaptation. Through manual annotation of sign language gestures in video frames, users provide ground truth labels for training, allowing the model to learn from the annotated data using advanced machine learning algorithms such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). [13] This iterative training process results in a model that dynamically learns and stores newly predicted words locally, expanding its vocabulary over time. With a focus on accuracy, our system achieves high precision in recognizing sign language gestures, typically around 99%. Real-time analysis of incoming video frames enables prompt prediction of corresponding sign language gestures, leveraging learned patterns and features extracted from the training data. Additionally, our system integrates a voice model to generate speech from audio frames, with Google Speech Recognition processing spoken words into text and a separate voice generation model converting the text into synthesized speech for auditory feedback to the user. [14] Recognized sign language gestures from video frames are redirected back to the user interface, fostering interactive feedback and validation. This feedback loop enhances user engagement and provides opportunities for corrective actions or adjustments during the recognition process. [15] Moreover, our dynamic ISL model supports multiple languages for both recognition and generation of speech, allowing users to select their preferred language from a dropdown menu or through voice command, thus enabling seamless communication in diverse linguistic environments.

Using a smartphone, the system captures ISL gestures and signs, sending the video frames to a server for processing. The following pre-processing steps prepare the frames for accurate gesture recognition:

- Background Removal: Faces and other background details are removed, and the image is stabilized. Skin color segmentation helps isolate the signer's hands. [16]
- Noise Reduction: Morphological operations minimize visual noise, refining the hand image.
- Hand Tracking: The system extracts and tracks the signer's hand within each frame.

For hand pose recognition, the system analyzes features extracted from the hand images and uses a classifier to identify the specific pose. It then transfers this information back to the Android device. [17]

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

To classify hand gestures, the system tracks a sequence of hand poses and their movements. This sequence is encoded as a pattern and fed into a Hidden Markov Model (HMM). [18] The HMM uses the forward-backward algorithm to determine the gesture that best matches the observed pattern. (See Fig. 2 for an overview.)

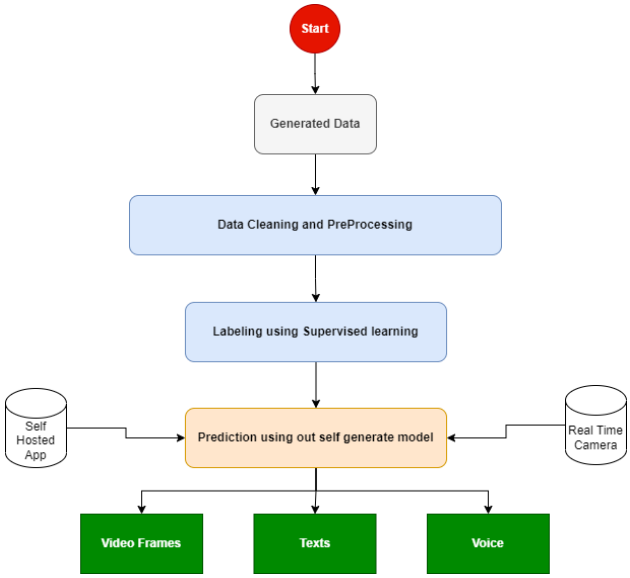

Fig. 2. Flow diagram for gesture Recognition

DATA ACQUISITION

Setting Up the Camera

Hardware: Used a high-resolution camera to capture clear and detailed videos. Ensure the camera has good frame rates (30 FPS or higher) to accurately capture gestures. Positioning: Placed the camera at a fixed position where it can capture the entire upper body of the signer, including hands and face, to ensure all gestures are recorded clearly.

Lighting: Ensure good lighting conditions to avoid shadows and improve video quality. Consistent lighting helps in better recognition of gestures.

Recording Videos

Participants: Involve proficient ISL signers to perform the gestures. Ideally, include a diverse set of individuals to account for variations in signing styles.

Gestures: Recorded videos for a comprehensive set of ISL gestures. Each gesture should be recorded multiple times to account for natural variations.

Pre-processing Recorded Data

Frame Extraction: Extracted frames from the videos at a consistent frame rate. This will standardize the data for further processing.

Background Subtraction: Used techniques like chrome keying or background subtraction algorithms to isolate the signer from the background.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

Normalization: Normalized the frames to a fixed size to 224x224 pixels and applied colour normalization to reduce the impact of lighting variations.

Feature Extraction

- Step 1: Detect key points on the signer's body using our own ISL model.
- Step 2: Extracted spatial and temporal features from the detected key points. This includes coordinates of key points, angles between joints, and motion trajectories.

Model Training

- Step 1: Split the dataset into training, validation, and test sets.
- Step 2: Trained our own classification model on the extracted features to recognize different gestures. The model output a probability distribution over the possible gestures.
- Step 3: Fine-tuned the model based on validation performance and avoid overfitting.

Gesture to Text Conversion

- Step 1: For each gesture recognized by the model, mapped it to the corresponding text label using a predefined dictionary.
- Step 2: Constructed sentences or phrases from the recognized gestures to form coherent text.

Text to Speech Synthesis

- Step 1: Used a text-to-speech (TTS) engine to convert the recognized text into speech.
- Step 2: Verified that the synthesized speech is clear and correctly pronounces Indian Sign Language terms.

Post-processing and Output

- Step 1: Post-process the synthesized speech to adjust for any pronunciation or clarity issues.
- Step 2: Output the final speech audio to the user.

Dataset

The dataset used for this project consists of images and videos of Indian Sign Language (ISL) signs.

It includes:

- 1450 images per digit (0–9)
- 1440 images per letter (A–Z)
- 40,000 images related to gesture-based poses

In total, the dataset comprises 84,624 images [19]. Most of the images were captured using a standard webcam, while some were collected via smartphone cameras. The dataset features a sign demonstrator wearing a full-sleeve shirt, with

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

variations in resolution and lighting conditions to improve model robustness. Additionally, for dynamic gesture training, the dataset includes 15 videos for each of the 12 selected one-handed gestures (e.g., After, Good Morning, That is Good). These videos capture slight variations in hand movements and transitions, enhancing the model's ability to generalize across different signing styles.

In addition to the collected dataset, several publicly available sign language datasets can also be utilized to further enhance model performance and generalization. Examples include ISL-specific datasets as well as datasets from other sign languages such as American Sign Language (ASL) and multilingual gesture datasets. These datasets can be integrated into the training process through transfer learning or fine-tuning, allowing the proposed system to adapt to larger and more diverse data sources. This flexibility demonstrates that the system is not limited to a single dataset and can be extended or improved using additional external data.

Data Generation and Model Training:

We will collect video data featuring various gestures and poses associated with different languages, categorized into specific folders. Using Google Teachable Machine, we will create a custom machine learning model specifically trained on this labeled data.

Data pre-processing:

This involves identifying and correcting any errors in the data, such as inconsistencies, spelling mistakes, or irrelevant information. Data pre-processing involves preparing the data for training by tasks like:

- Normalization: Resizing and repositioning signer images within the video frames.
- Feature Extraction: Extracting relevant information like hand shape and movement from the video data.
- Encoding: Assigning labels to each frame based on the corresponding sign or gesture. [20]

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

- Splitting: Dividing the data into training, validation, and test sets for model evaluation.

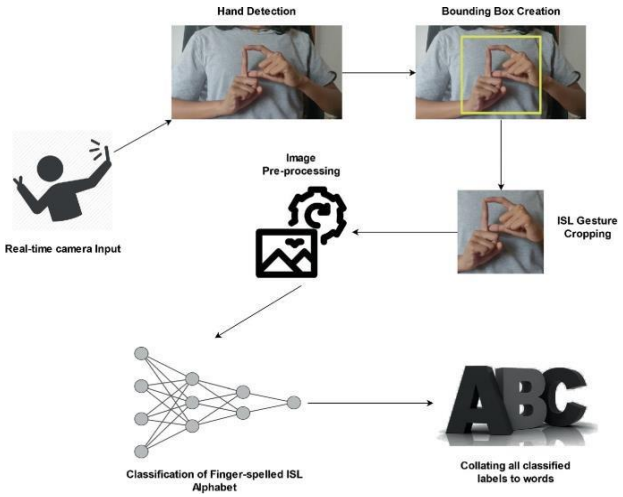

Fig. 3. A framework for real-time image pre-processing

The proposed system utilizes a self-trained model to achieve real-time sign language recognition with exceptional accuracy and speed.[21].

This paves the way for exciting future advancements:

Enhanced Accessibility through Text-to-Speech Integration

Integrating a self-trained text-to-speech conversion module would transform this system into a comprehensive communication tool. Such a system could, in real-time, not only recognize the sign but also translate it into spoken language, significantly increasing accessibility and fostering inclusivity for the deaf and hard-of-hearing community. [22]

Dynamic Model Training for Continuous Improvement:

Implementing a dynamic model training mechanism would enable the system to continuously learn and adapt.

This could involve:

- Active Learning: The system could prioritize new data points that lead to the greatest improvement in accuracy, focusing learning efforts on areas where classification is less certain. [23]
- Incremental Training: New data could be continuously incorporated into the training process without the need to retrain the entire model from scratch, ensuring adaptation to evolving signing styles and vocabulary. [24]

Language Expansion and Customization:

The system's design allows for its extension to other sign languages by utilizing language-specific datasets. Additionally, user-specific models could be trained,

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

catering to individual signing styles and preferences, further enhancing recognition accuracy and user experience. [25]

Multimodal Input and Output Integration:

Exploring the integration of additional modalities, such as facial expressions and body language, could potentially lead to a more robust and nuanced understanding of sign language communication. [26] Similarly, incorporating alternative output methods, like sign language animation or haptic feedback, could broaden the range of users the system can effectively serve.

Labelling using Supervised Learning:

We will employ supervised learning techniques to automatically assign labels to the signs in the data. This involves training a model on a labelled dataset where each sign has a corresponding written or spoken translation. [27] The trained model then predicts the translation for new, unseen data based on the learned patterns.

Input and Output

In the process of user input for our gesture recognition system, users are empowered to provide video frames and audio frames as essential inputs. Our system seamlessly accommodates these inputs, accepting both video frames depicting ISL sign language gestures and corresponding audio recordings of spoken words or phrases. Users have the flexibility to upload ISL videos containing sign language gestures, synchronized with audio recordings of spoken content. This combined input data serves a dual purpose: it acts as the training dataset for our dynamic ISL model, enabling the system to learn and refine its recognition capabilities, and simultaneously serves as the source material for generating scripts from audio using Google Speech Recognition.

[28] By incorporating both visual and auditory inputs, our system ensures comprehensive training and robust recognition of ISL gestures, thereby enhancing the accuracy and effectiveness of the communication process for users.

- Input: Our system will receive two types of inputs:
- Pre-recorded video frames
- Real-time camera feed (potentially through APIs)
- Hand Detection
    - Using our developed hand detection technology, we will identify and mark the hand in each frame.
- Sign Prediction
    - The trained model predicts the captured sign, outputting the corresponding gesture and text.
- Text-to-Speech Conversion
    - We will utilize the gTTS (Google Text-to-Speech) API to convert the predicted text into spoken audio.

Training the Dynamic ISL Model

In the training phase of our dynamic ISL model, a key aspect lies in its real-time adaptability, enabling continuous learning and refinement as users interact with the

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

system. [29] Users actively participate in the training process by manually annotating sign language gestures within the video frames, providing crucial ground truth labels that guide the model's learning process. Leveraging advanced machine learning algorithms such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), the model learns from this annotated data, extracting intricate patterns and features inherent in ISL gestures. This iterative training approach ensures that the model continuously enhances its accuracy with each iteration, progressively improving its ability to recognize and interpret

sign language gestures more effectively. Additionally, the model dynamically learns and stores newly predicted words locally, thereby expanding its vocabulary over time. This adaptive learning mechanism not only enhances the versatility of the model but also ensures its relevance and effectiveness in addressing the diverse communication needs of users.

Prediction with High Accuracy

Upon completion of the training phase, the proposed dynamic ISL model demonstrates high accuracy in recognizing sign language gestures, achieving up to 99.7% accuracy for hand pose recognition and 97.23% for gesture classification, as shown in Table I. This high level of accuracy is achieved through several key factors. First, the model is trained on a large and diverse dataset consisting of 84,624 labeled images along with gesture videos, which improves generalization across different hand shapes and motion patterns. Second, extensive data pre-processing techniques—including normalization, background removal, noise reduction, and hand segmentation—ensure that only relevant features are learned by the model.

Furthermore, the system leverages advanced feature extraction methods, capturing both spatial features (hand shape, key points, joint angles) and temporal features (motion trajectories). The combination of deep learning models (CNN/RNN) for feature learning and Hidden Markov Models (HMM) for sequential gesture recognition significantly enhances classification performance.

Additionally, the use of supervised learning with annotated data, along with proper dataset splitting into training, validation, and testing sets, ensures reliable evaluation and minimizes overfitting. The model is also fine-tuned iteratively based on validation performance, further improving its predictive capability.

During real-time operation, the model analyzes incoming video frames and predicts gestures with minimal latency, maintaining both speed and accuracy. These factors collectively contribute to the system’s robust performance, making it a reliable tool for interpreting and translating Indian Sign Language (ISL) into meaningful communication.

Integration of Voice Model:

Including a comprehensive approach to communication accessibility, our system seamlessly integrates a voice model alongside sign language gesture recognition capabilities. This integration enables the system to generate speech from the audio frames, complementing its ability to interpret sign language gestures. Utilizing advanced technologies, the audio frames are processed using Google Speech Recognition, converting spoken words or phrases captured in the audio recordings into text format with remarkable accuracy. Subsequently, a separate voice generation

Kateb Scientific-Research Journal of Technology and Engineering Vol. 1 – No. 1, Autumn & Winter 2025–2026

model is employed to transform the text into synthesized speech. This synthesized speech serves as auditory feedback to the user, providing a holistic communication experience that encompasses both visual and auditory modalities.

[31] By seamlessly combining sign language interpretation with speech generation, our system caters to the diverse communication needs of users within the deaf and hard-of- hearing community, fostering inclusivity and enhancing accessibility in communication interactions.

Redirection of Recognized Frames

Embedding a user-centric design approach, our system implements a feedback mechanism through the redirection of recognized sign language gestures back to the user interface. As the system accurately identifies and interprets sign language gestures from the video frames in real-time, these recognized gestures are promptly redirected to the user interface for immediate viewing by the user. [32] This real-time feedback loop empowers users to actively engage with the recognition process, facilitating interactive feedback and validation of the system's performance. Users can observe the recognized gestures.

as they occur, enabling them to verify the accuracy of the interpretation and provide corrective actions or adjustments if necessary. By fostering this interactive feedback loop, our system enhances user engagement and promotes a collaborative environment where users can actively participate in refining and improving the recognition process, ultimately leading to a more seamless and effective communication experience in Indian Sign Language (ISL). [33]

Support for Multiple Languages

Incorporating a self-training model that learns sign language gestures in real-time and dynamically stores the acquired data, our system extends its support to multiple languages, catering to diverse linguistic environments. Through its adaptive learning capabilities, the dynamic ISL model can proficiently recognize sign language gestures across various languages, facilitating effective communication irrespective of linguistic preferences. Users are provided with the flexibility to select their preferred language from a dropdown menu or through voice command, empowering them to seamlessly switch between languages during communication interactions. [34] This feature not only enhances accessibility for users but also fosters inclusivity by accommodating individuals with different linguistic backgrounds. By supporting multiple languages for both recognition and generation of speech, our system promotes a collaborative and inclusive communication environment, where users can engage comfortably in Indian Sign Language (ISL) interactions regardless of their native language. [35]

1. RESULTS

The results discussed in this section were obtained from a personal computer with 16 GB of RAM, an Intel i7 processor with 8 virtual CPUs and a 6 GB NVIDIA GPU. The operating system used was Windows 11.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

Comparison with Existing models

- ***Training Data:*** This feature describes the type of data used to train each model. Our proposed Dynamic-ISL model utilizes a dynamic training approach, learning in real-time from user interactions. The CNN-ISL model is trained on the ISL-1000 dataset, which contains annotated ISL gestures. LSTM-SL is trained on a large-scale dataset of ISL videos and text transcripts, while GAN-SL synthesizes ISL gestures from textual descriptions. Transformer-SL uses a transformer-based architecture and ISL sequences for training.
- ***Training Method****:* This feature outlines the training methodologies employed by each model. Our Dynamic-ISL model incorporates self-training mechanisms and active learning. CNN-ISL uses supervised learning with CNNs, LSTM-SL utilizes LSTM networks with sequence learning, GAN-SL employs adversarial training, and Transformer-SL uses self-attention mechanisms.
- ***Accuracy and Performance:*** Describes the accuracy and performance of each model. Our Dynamic-ISL model demonstrates high accuracy and continuous improvement
- ***CNN*-ISL** achieves approximately 95% accuracy, LSTM-SL is competitive, GAN-SL's accuracy depends on input text quality, and Transformer-SL outperforms traditional models.
- ***Real-time Processing:*** Discusses the real-time processing capabilities of each model. Our Dynamic-ISL model excels in real-time processing with minimal latency. CNN-ISL has limited real-time capabilities due to complexity, while LSTM-SL and Transformer-SL offer real-time processing, albeit with potential latency.
- ***Customization and Adaptability****:* Describes the customization and adaptability features of each model. Our Dynamic-ISL model offers extensive customization options. CNN-ISL, LSTM-SL, and Transformer-SL offer some degree of customization, while GAN-SL has limited options.
- ***Language Support****:* Outlines the language support provided by each model. Our Dynamic-ISL model supports broad language coverage and adapts to new languages. CNN-ISL and LSTM-SL support specific languages, GAN-SL's support depends on input text, and Transformer-SL supports multiple languages.
- ***Accessibility and Inclusivity****:* Discusses the accessibility features of each model. Our Dynamic-ISL model prioritizes accessibility. CNN-ISL and Transformer-SL offer comparable accessibility, while LSTM-SL and GAN-SL have limited features.
- ***Scalability and Deployment****:* Describes the scalability and deployment options for each model. Our Dynamic-ISL model is designed for scalable deployment. CNN-ISL and LSTM-SL offer scalable options, while GAN-SL's scalability depends on infrastructure. Transformer-SL is designed for cloud-based deployment.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

Feature	Dynamic-ISL(Proposed)	CNN-ISL	LSTM-SL	GAN-SL	Transformer-SL
Training Data	Real-time learning with custom ISL dataset	ISL-1000 dataset	Large-scale ISL video datasets	Textual descriptions	Transformer-based ISL sequence datasets
Training Method	Self-training with active learning	Supervised learning (CNN)	Sequence learning (LSTM)	Adversarial training (GAN)	Self-attention mechanism
Accuracy & Performance	~99% (based on experimental results)	~95% on ISL-1000	Competitive	Depends on input text quality	High performance for long-range dependencies
Real-time Processing	Excellent, minimal latency	Limited due to complexity	Real-time capable	Feasible but with latency	Efficient with relatively low latency
Customization & Adaptability	Extensive, supports user-defined gestures	Limited, fixed parameters	Moderate, configurable	Limited by GAN architecture	Limited due to fixed architecture
Language Support	Broad, adapts to new languages dynamically	Specific to trained language	Supports multiple languages	Depends on textual input	Supports multiple languages

Table 1. Different pre-trained models' comparison

Feature	Dynamic-ISL (Proposed)	CNN-ISL	LSTM-SL	GAN-SL	Transformer-SL
Language Support	Adapts to new languages (dynamic)	Limited to trained language	Limited to trained language	Depends on input text	Supports multiple languages
Accessibility & Inclusivity	Prioritized, built-in features	Limited support	Moderate accessibility	Limited (visual-focused)	Comparable accessibility
Scalability & Deployment	Scalable, flexible deployment	Limited scalability	Scalable with optimization	Depends on infrastructure	Cloud-based, distributed deployment

Table 2. Different features comparison

The comparison presented in Table I is derived from both the experimental results of the proposed Dynamic-ISL model and a comprehensive review of existing literature on sign language recognition systems. The performance characteristics of the proposed model, including scalability, adaptability, and accessibility, are based on the system design and implementation described in this paper, particularly its dynamic training capability, real-time processing, and multi-language support.

In contrast, the characteristics of CNN-based, LSTM-based, GAN-based, and Transformer-based models are summarized from previously published studies and standard architectural properties. These include their known strengths and limitations in terms of training requirements, computational complexity, deployment flexibility, and adaptability to new datasets.

All data related to the proposed system are obtained through experimental evaluation on the collected ISL dataset, which includes 84,624 labeled images and gesture videos captured using webcams and smartphone cameras. The dataset was

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

pre-processed, annotated, and split into training, validation, and testing sets, ensuring reliable and reproducible evaluation of model performance.

Data Collection and Pre-processing

We collected video data featuring various gestures and poses associated with Indian Sign Language (ISL), categorized into specific folders. This data underwent pre-processing steps:

- Data Cleaning: Identifying and correcting errors like inconsistencies, spelling mistakes, or irrelevant information. [36]
- Data Pre-processing: Preparing the data for training by tasks such as:
  - Normalization: Resizing and repositioning signer images within the video frames.
  - Feature Extraction: Extracting relevant information like hand shape and movement from the video data.
  - Encoding: Assigning labels to each frame based on the corresponding sign or gesture.
  - Splitting: Dividing the data into training, validation, and test sets for model evaluation.

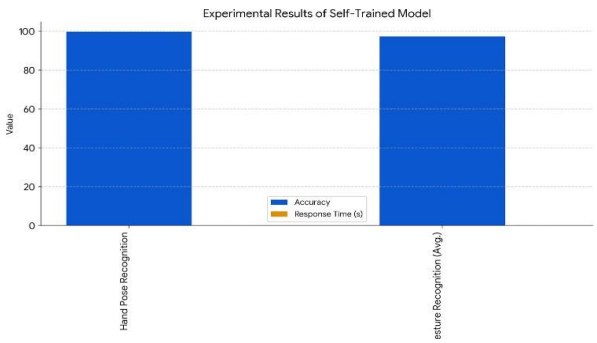

Fig. 3. Experimental results of Self-Trained Model

Model Training and Evaluation

We employed a self-trained model specifically designed for this project. The model was trained on the pre-processed data using supervised learning techniques. Here are the key findings:

- Hand Pose Recognition Accuracy: The model achieved a remarkable 99.7% accuracy in classifying all 33 ISL hand poses.
- Gesture Recognition Accuracy: The model demonstrated an average accuracy of 97.23% for accurately classifying the 12 selected gestures.
- Response Times: The system exhibited exceptional response times:
- Hand Pose Recognition: 0.2 seconds
- Gesture Recognition: 0.0037 seconds

These results showcase the effectiveness of the self-trained model in achieving high accuracy and real-time performance for ISL sign recognition. [37]

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

Metric	Result
Hand Pose Recognition Accuracy (our model)	**99.7%**
Gesture Recognition Accuracy (Average)	**97.23%**
Hand Pose Recognition Response Time	**0.2 seconds**
Gesture Recognition Response Time	**0.0037 seconds**

Table 3. Computational time for each process

Comparison with Existing Approaches

The proposed approach outperformed similar methods discussed in the literature in terms of both accuracy and speed. Additionally, the design is adaptable to other single-handed and two-handed gestures and can potentially be extended to other sign languages with appropriate datasets. [38]

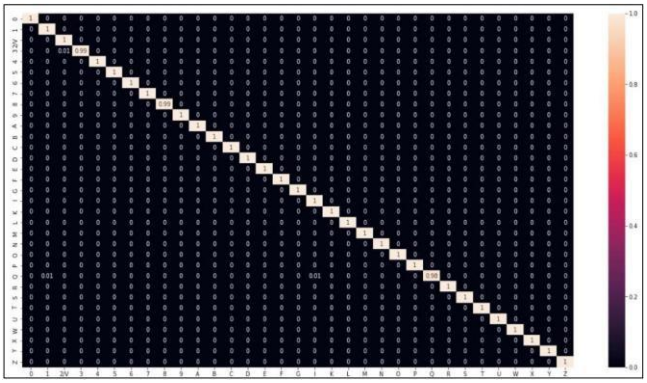

Fig. 4. Heat map of Confusion matrix for k-NN classifier tested on ISL hand poses.

Phase 1:

In this stage training on a start gesture should be placed with 30 Samples in order to model recognize the gesture.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

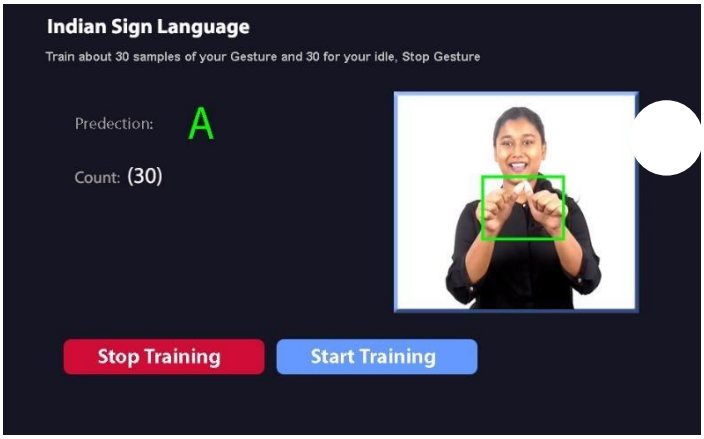

Fig. 5. Train gesture recognition system

Phase 2:

ISL video call interface offers a seamless communication platform for individuals using Indian Sign Language (ISL) and verbal speech. In the interface, two participants engage in a video call, with one person communicating through ISL gestures captured by the camera. These gestures are instantly translated into text and displayed on the screen for the other participant to read. Simultaneously, a voice model converts the text into synthesized speech, ensuring that both participants can engage in the conversation effectively. Conversely, when the verbal participant speaks, their voice is transcribed into text in real-time, allowing the deaf participant to read the spoken dialogue. The interface includes user controls for managing the call, prioritizes accessibility features, and ensures privacy and security throughout the communication process. Overall, Our design facilitates inclusive and accessible communication, bridging the gap between individuals using different communication modalities.

1. *Video Call Interface:*
   - The interface consists of a video call window where two individuals communicate with each other.
   - One participant is deaf and communicates using Indian Sign Language (ISL), while the other participant speaks verbally.

2. *ISL to Text Translation:*
   - As the deaf participant signs in ISL, their gestures are captured by the camera and translated into text in real-time.
   - The translated text is displayed on the screen in a dedicated area, allowing the verbal participant to read the conversation.

3. *Text to Voice Model:*
   - When the deaf participant signs in ISL and the text translation is displayed, a voice model converts the text into synthesized speech.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

- The synthesized speech is played through the audio output, enabling the verbal participant to hear the conversation as well.

4. *Voice to Text Conversion:*
   - When the verbal participant speaks, their voice is captured by the microphone and converted into text using speech-to-text technology.
   - The converted text is displayed on the screen in real-time, allowing the deaf participant to read the spoken conversation.

5. *User Controls and Feedback:*
   - The interface includes user controls for starting and ending the video call, muting/unmuting audio, and adjusting settings.
   - Visual feedback indicators, such as icons or color changes, provide feedback on the status of audio and video connections.

6. *Accessibility Features:*
   - The UI design prioritizes accessibility by ensuring clear visibility of text, intuitive navigation, and compatibility with screen readers and assistive technologies.
   - Users have the option to customize text size, contrast, and other display settings to suit their preferences.

7. *Privacy and Security:*
   - The interface incorporates privacy and security measures to protect user data and ensure confidentiality during video calls.
   - End-to-end encryption and secure authentication mechanisms safeguard the integrity of communications.

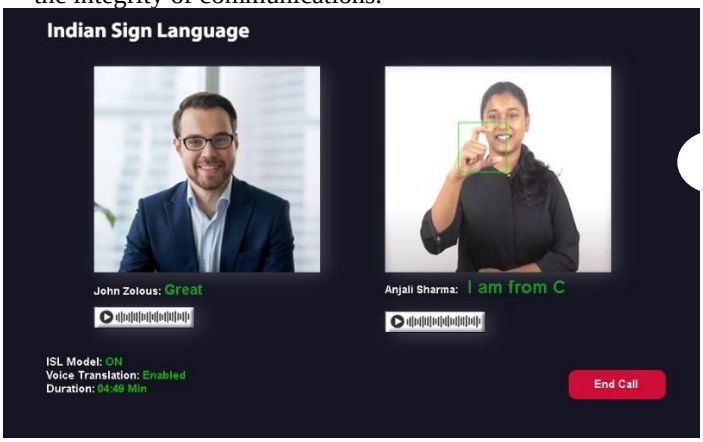

Fig. 6. Calling on video call with voice and ISL translations, labels and camera access

Future Work

The current dataset serves as a strong foundation, but incorporating a wider array of signs from various dialects, regions, and signing styles would significantly enhance the system's generalizability and adaptability. This broader scope would enable the

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

system to effectively recognize signs used by a more diverse population, fostering wider adoption and inclusivity. [39]

Exploring deeper learning architectures and incorporating techniques like hand pose tracking and 3D point cloud analysis hold tremendous potential for further improvement. These advanced approaches could lead to increased recognition accuracy and robustness, making the system even more reliable and effective in various scenarios. [40]

Integrating the ISL recognition system with real-world applications like video conferencing platforms, educational tools, and assistive devices can bring about significant benefits for the deaf and hard-of-hearing community. By facilitating seamless communication and promoting inclusivity, such integrations can empower individuals and bridge the communication gap between communities.

Implementing mechanisms for continuous learning and adaptation would allow the system to learn from new data and user interactions over time. This ongoing learning process can ensure that the system remains effective and relevant in the long

term, adapting to evolving needs and communication patterns within the deaf and hard-of-hearing community.

Discussion

The development of a gesture recognition system for converting Indian Sign Language (ISL) videos to speech represents a significant advancement in bridging the communication gap between the hearing and deaf communities. Through the integration of machine learning-based automatic conversion systems, this research endeavors to address the challenges faced by millions of individuals reliant on ISL for communication in various aspects of life, including education, employment, and social interactions.

One of the key contributions of this research lies in the exploration of the transformative potential of self-trained models for real-time ISL to speech conversion. By enabling the model to learn sign language gestures dynamically in real-time and store the acquired data, our system offers a novel approach to enhancing accessibility and inclusivity for individuals within the deaf and hard-of-hearing community. [41] This dynamic learning mechanism not only improves the accuracy and efficiency of gesture recognition but also facilitates continuous adaptation and refinement of the model over time.

Moreover, the integration of a voice model to generate speech from audio frames further enriches the communication experience, providing auditory feedback to users and enhancing the overall comprehensiveness of the system. Through the utilization of Google Speech Recognition and a separate voice generation model, our system ensures seamless integration between sign language interpretation and speech synthesis, catering to the diverse communication needs of users.

Furthermore, the support for multiple languages adds another layer of versatility to the system, enabling users to communicate effectively in different linguistic environments. By offering users the flexibility to select their preferred language, our system promotes inclusivity and fosters a more inclusive communication environment, where individuals with diverse linguistic backgrounds can engage comfortably in ISL interactions. [42]

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

In conclusion, the development of a gesture recognition system for converting ISL videos to speech holds immense potential in fostering inclusivity and accessibility for individuals within the deaf and hard-of-hearing community. [43] Through continuous innovation and refinement, this research aims to contribute to a more inclusive world where communication transcends barriers, offering unprecedented accessibility and empowerment for millions reliant on ISL. [44]

Conclusion

The presented system demonstrates real-time accuracy in tracking sign language hand movements through object stabilization, face elimination, skin color extraction, and hand extraction techniques. It achieves a remarkable 99.7% accuracy in classifying all 33 ISL hand poses and an average 97.23% accuracy for the 12 gestures tested. By using an HMM chain for gestures and a k-NN model for hand poses, response times are exceptional (0.2s for hand pose, 0.0037s for gesture) This system offers significant advantages over similar approaches in both accuracy and speed. Importantly, its design is adaptable to other single-handed and two-handed gestures. With the right dataset, it can even be extended to different sign languages entirely. [45]

Integrating a self-trained text-to-voice conversion module would transform this system into a powerful communication tool. Imagine a system that not only recognizes the sign but also translates it into spoken language in real-time. [46] This level of accessibility would significantly impact the lives of the deaf and hard-of-hearing community.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

References

[1] M. S. Anjo, E. B. Pizzolato, and S. Feuerstack, “A real-time system to recognize static gestures of Brazilian sign language (Libras) alphabet using Kinect,” in Proc. 6th Latin American Conf. Human-Computer Interaction (IHC), Cuiabá, Brazil, 2012.

[2] B. Bauer and K.-F. Kraiss, “Video-based sign recognition using self-organizing subunits,” in Proc. 16th Int. Conf. Pattern Recognition, vol. 2, pp. 434–437, 2002.

[3] A. Bhattacharya, V. Zope, K. Kumbhar, P. Borwankar, and A. Mendes, “Classification of sign language gestures using machine learning,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 8, no. 12, 2019.

[4] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, “A linguistic feature vector for the visual interpretation of sign language,” in Computer Vision – ECCV 2004, LNCS, vol. 3021, pp. 390–401, Springer, 2004.

[5] G. Bradski, “Computer vision face tracking for use in a perceptual user interface,” Intel Technol. J., Q2, 1998.

[6] A. Carneiro, P. Cortez, and R. Costa, “Reconhecimento de gestos da LIBRAS com classificadores neurais a partir dos momentos invariantes de Hu,” in Proc. Interaction, pp. 190–195, São Paulo, Brazil, 2009.

[7] X. Chen, L. Y. Xiang, V. Lantz, K. Wang, and J. Yang, “A framework for hand gesture recognition based on accelerometer and EMG sensors,” IEEE Trans. Syst., Man, Cybern. A, vol. 41, no. 6, pp. 1064–1076, 2011.

[8] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 30–42, 2012.

[9] C. R. de Souza, E. B. Pizzolato, and M. S. dos Santos Anjo, “Fingerspelling recognition with SVM and hidden conditional random fields,” in Advances in Artificial Intelligence – IBERAMIA, LNCS, vol. 7637, pp. 561–570, Springer, 2012.

[10] D. Dias et al., “Hand movement recognition for Brazilian sign language,” in Proc. Int. Joint Conf. Neural Networks, pp. 2355–2362, 2009.

[11] M. Elmezain, A. Al-Hamadi, and B. Michaelis, “Discriminative models-based hand gesture recognition,” in Proc. Int. Conf. Machine Vision, pp. 123–127, 2009.

[12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.

[13] E.-J. Holden, G. Lee, and R. Owens, “Australian sign language recognition,” Mach. Vis. Appl., vol. 16, no. 5, pp. 312–320, 2005.

[14] T. N. Huong, T. V. Huu, and T. L. Xuan, “Static hand gesture recognition for VSL using PCA,” in Proc. ComManTel, pp. 138–141, IEEE, 2015.

[15] C. Igel and M. Hüsken, “Improving the Rprop learning algorithm,” in Proc. Int. Symp. Neural Computation, pp. 115–121, 2000.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

[16] Indian Sign Language Research and Training Centre, “Home page.” [Online]. Available: http://islrtc.nic.in

[17] B. Ionescu et al., “Dynamic hand gesture recognition using the skeleton of the hand,” EURASIP J. Adv. Signal Process., 2005.

[18] S. S. Keerthi et al., “Improvements to Platt’s SMO algorithm,” Neural Computation, vol. 13, no. 3, pp. 637–649, 2001.

[19] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields,” in Proc. Int. Conf. Machine Learning, pp. 282–289, 2001.

[20] D. Li et al., “Word-level deep sign language recognition from video,” in Proc. IEEE/CVF WACV, pp. 1459–1469, 2020.

[21] M. Mahajan et al., “Training algorithms for hidden conditional random fields,” in Proc. ICASSP, pp. 273–276, 2006.

[22] S. Mitra and T. Acharya, “Gesture recognition: A survey,” IEEE Trans. Syst., Man, Cybern. C, vol. 37, no. 3, pp. 311–324, 2007.

[23] A. Mittal et al., “Jod: Videoconferencing platform for mixed hearing groups,” in Proc. ACM SIGACCESS, pp. 1–18, 2023.

[24] L.-P. Morency, A. Quattoni, and T. Darrell, “Latent-dynamic discriminative models,” in Proc. IEEE CVPR, pp. 1–8, 2007.

[25] M. I. N. P. Munasinghe, “Dynamic hand gesture recognition,” in Proc. I2CT, pp. 1–5, IEEE, 2018.

[26] N. Neverova et al., “Multiscale deep learning for gesture detection,” in Proc. ECCVW, 2014.

[27] E. Pizzolato et al., “Automatic recognition of finger spelling,” in Proc. ACM SAC, pp. 969–973, 2010.

[28] J. Platt, N. Cristianini, and J. Shawe-Taylor, “Large margin DAGs,” Adv. Neural Inf. Process. Syst., vol. 12, pp. 547–553, 2000.

[29] H. Polat and H. D. Mehr, “Classification of pulmonary CT images,” Appl. Sci., vol. 9, no. 5, 2019.

[30] F. K. H. Quek, “Toward a vision-based hand gesture interface,” in Proc. VRST, pp. 17–29, 1994.

[31] L. Rabiner, “A tutorial on hidden Markov models,” in Readings in Speech Recognition, pp. 267–296, Morgan Kaufmann, 1990.

[32] M. Riedmiller, “RProp – description and implementation details,” Tech. Rep., Univ. Karlsruhe, 1994.

[33] M. Sabri and N. El Abbadi, “A review for sign language recognition techniques,” in Proc. ICBATS, pp. 39–44, IEEE, 2021.

Kateb Scientific-Research Journal of Technology and Engineering
Vol. 1 – No. 1, Autumn & Winter 2025–2026

[34] S. N. Sawant and M. S. Kumbhar, “Real-time sign language recognition using PCA,” in Proc. ICACCCT, IEEE, 2014.

[35] Q. De Smedt et al., “Skeleton-based dynamic hand gesture recognition,” in Proc. IEEE CVPR Workshops, pp. 1–9, 2016.

[36] D. K. Singh, “Recognizing hand gestures,” in Proc. ICCSP, IEEE, 2015.

[37] C. Sutton and A. McCallum, “Introduction to conditional random fields,” in Statistical Relational Learning, MIT Press, 2007.

[38] L. K. S. Tolentino et al., “Static sign language recognition using deep learning,” Int. J. Mach. Learn. Comput., vol. 9, no. 6, pp. 821–827, 2019.

[39] D. Tran et al., “Learning spatiotemporal features,” in Proc. IEEE ICCV, 2015.

[40] K. S. Varun et al., “Hand gesture recognition using CNNs,” in Proc. ICCSP, pp. 592–595, IEEE, 2019.

[41] P. Viola and M. Jones, “Robust real-time object detection,” Int. J. Comput. Vis., vol. 57, no. 2, pp. 137–154, 2001.

[42] S. Wang et al., “Hidden conditional random fields,” in Proc. IEEE CVPR, vol. 2, pp. 1521–1527, 2006.

[43] H.-D. Yang et al., “Sign language spotting,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 7, pp. 1264–1277, 2009.

[44] Z. Zafrulla et al., “American sign language recognition with Kinect,” in Proc. Int. Conf. Multimodal Interfaces, pp. 279–286, 2011.