# Speech impaired people communicate to Real-Time Sign Language Detection with Deep Learning

Kowsalya G[1], Reenadevi N[2], Sajetha K[3], Prasanth T[4], Sowmiya M[5]
[1,2,3,4,5]*Department of Computer Science and Engineering The kavery Engineering College, Mecheri, Salem Tamilnadu, India -636 453*

***Abstract*—People who have problems with their speech frequently face enormous challenges to communicate due to barriers that restrict their ability to communicate in real-time, get important help, and engage in social activities. This project addresses these difficulties by using a real-time hand gesture recognition system based on CNN, which will allow gestures to be converted immediately into both text and audio forms. The system will also recognize a wide variety of gestures and signs, including emergency signs, so that critical communication may occur quickly during emergencies. To assist in making this process available to as many people as possible, the platform will produce multiple outputs in the three major languages spoken in India, making the project useful for users across different language groups. Additionally, the interface displays the corresponding text-based sign symbols, assisting users who are either learning how to sign or interacting with others who do not know how to sign. By combining cutting-edge machine-learning techniques with attention to user-friendly visual and auditory feedback, this system will bring individuals with speech impairments and their communities together socially, thus increasing social interaction, education, and access to public services. In addition, this innovative approach will give users more independence and confidence while increasing the visibility and understanding of sign language, thereby creating a more inclusive and accessible environment for people with speech and communication discrepancies.**

***Index Terms*—Speech impairment, Sign language recognition, Hand gesture recognition, Convolutional Neural Networks (CNN), Real-time translation, Multilingual support, Assistive technology, Inclusive communication.**

## I. INTRODUCTION

The field of sign language recognition research to facilitate inclusive communication amongst those who are hearing and/or speech impaired has gained prominence in the last few years. The SignEval 2025 Challenge held at ICCV demonstrated to researchers that multimodal approaches, benchmarking datasets at scale, and assessing the robustness of sign recognition systems using data collected in real-world conditions are of growing importance to the development of sign language recognition systems [1]. The challenge highlighted several key milestones for future research in this field, including continuous sign recognition, domain adaptation, and signer-independent evaluation, all of which mirror the practical requirements for deploying these systems in the real world. In addition, many recent advances in deep learning-based pathways for automatic processing of sign language suggest a shift from manually crafted features to end-to-end neural architecture that models spatial-temporal dependencies effectively [2]. This research demonstrates that AI-driven systems now outperform traditional rule-based systems in terms of scalability and accuracy in real-time applications.

Deep computer vision methods that use artificial intelligence have also enhanced sign language identification systems by using the following methods: convolutional neural networks (CNNs) as well as transfer learning models [3]. These systems are able to obtain discriminative spatial features of hand gestures, face expressions and body posture which will improve classification accuracy in a variety of environmental conditions. In addition, research into systems that convert automatic spoken

language into signed languages has shown that interdisciplinary research can be performed by integrating speech recognition, natural language processing and gestures models of generation [4]. Therefore, these integrated systems not only create a means for recognizing signs, but they also create a means of bilaterally communicating between the hearing and deaf communities and increasing accessibility to sign language across digital platforms.

Virtual reality technology is also being studied to enhance sign language learning and use beyond recognition or translation. Virtual environments create an interactive place for users to learn to use gestures with immediate feedback, showing how traditional classroom methods are limited in providing a place for hands-on work with other signers. These advances suggest a shift away from isolated systems to comprehensive systems for communicating that leverage deep learning, multimodal fusion and human-computer interaction framework. All five studies show the need for sign language systems that are robust, scalable and intelligent to be effective in real-world situations that are constantly changing.

As a result of these new developments, deep learning research today is looking to create real time recognition models based on CNN architectures combined with optimized data sets, and adaptive preprocessing techniques. Modern systems attempt few of the challenges faced by a signer level of variability; there are also lighting conditions and continuous gesture segmentation. Overall, it is felt that the three areas (computer vision, multimodal AI, and accessibility-driven design) will come together to define a new era of intelligent sign language technologies which will create inclusive communication opportunities around the world. [1][2][3][4][5]

## II. LITERATURE REVIEW

Beyond day-to-day interactions, sign language has enormous significance in multiple areas of life, including healthcare. Reports from professionals indicate that gaps exist between medical professionals who need to converse with a deaf patient. There is an urgent need for the development of automatic tools for communicating with deaf patients by way of translation and recognition in clinical settings [6]. A number of researchers have also suggested using hybrid deep-learning models, which combine Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN), to improve the accuracy of recognizing people who have hearing and speaking disabilities [7]. Additionally, research has documented sociocultural impediments regarding the acceptance of national sign languages and how they were exacerbated by the global COVID-19 pandemic, pointing out the need for platforms that provide access to digital means for inclusive communication [8]. Recent literature reviews regarding Continuous Sign Language Recognition (CSLR) have identified the need for temporal modeling processes to fully represent and analyze dynamic gesture sequences using LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), and Transformer models [9].

Examples of web-based, scalable systems which connect Deaf and hearing individuals more easily are found in Auslanweb's practical implementations on cloud-based platforms (see [10]). Previous methods of recognition of human movement used either segmented skin colours or geometrical templates to define a function; however, these were not linear functions capable of expressing motion in the same way that Hidden Markov Models (HMMs) are [11]. Probabilistic techniques such as an HMM can model a sequence of numbers from a video stream using a mathematically-based model and therefore support dynamic recognition of hand gestures [12]. Another application area for HMM is facial expression recognition, in which the use of HMM with various feature descriptors (e.g Zernike moments) has played an important role in multimodal gesture analysis systems [13]. The combination of early work on the model-based and feature-based approaches have helped establish current systems using deep-learning based approaches.

The use of Texture and Shape-Based Feature Extraction techniques has been explored in order to improve the accuracy of gesture classification for content-based image retrieval (CBIR) systems [14]. Multi-feature Fusion techniques were employed to combine broad-ranging colour, motion, and contour descriptors with template matching strategies in order to establish greater robustness for controlled environments [15]. The use of depth (3D motion

data) for gesture recognition through Dynamic Time Warping (DTW) has improved dynamic sequence alignment [16]. These methods have been found to provide moderate levels of accuracy; however, they are restricted by their high sensitivity to background noise/variations in lighting as well as their relative computational inefficiencies. Consequently, real-time systems have limited scalability due to these constraints.

Recurrent neural networks (RNNs) and two-stream architectures were among the first successful methods for achieving large-scale continuous gesture recognition [17]. Since they use both image data for spatial features and motion data for capturing motion dynamics, the temporal consistency of their gesture recognition output has greatly improved. In addition, because unsupervised gesture segmentation techniques use motion detection to automatically find gesture boundaries within a continuous stream, they provide another way of enhancing real-time efficiency [18]. The use of multi-scale deep learning frameworks for gesture detection and localization also demonstrated the usefulness of hierarchical feature extraction to detect complex objects in images [19]. Compared to traditional hand-crafted methods, these types of methods greatly increased the accuracy of the localization process.

The field of sign language recognition has seen major changes as researchers move away from creating manual feature-based models to developing deep neural network architectures, showing clear signs of progression. Past research provided an early foundation of algorithms for segmentation, feature extraction, and sequential modelling [11]– [16] before moving towards a future where researchers created end-to-end deep learning architectures that process multiple modalities and recognize continuous (i.e., non-discrete) signs [17]– [20]. There are still several challenges remaining, including lack of diversity in available datasets, signer independence, computational efficiency, and real-time delivery. To overcome these challenges, researchers will need to combine state-of-the-art CNN-based spatial modeling techniques with leading-edge temporal learning techniques in order to create scalable and inclusive communication systems suitable for use in real-world environments.

## III. METHODOLOGY

The real-time sign language recognition solution consists of a structured workflow. This allows for efficient detection, translation, and accurate recognition of each gesture involved in sign language. The system begins by capturing video input (screenshots) through a camera attached (a webcam or mobile device). Once captured, these frames are pre-processed so that they can then be processed into images that will ultimately be fed into the CNN model for recognition. Pre-processing steps include the following: resizing to a standard size - (x,y), normalizing pixel values (0-255), reducing noise from the images, utilizing augmentation methods (rotation, flipping, changing brightness), etc.). These actions will increase the CNN model's robustness against varying light sources, different backgrounds, and different orientations of users' hands. After the sign language gesture is captured, processed, labeled according to the specific sign language gesture it corresponds to, and stored in a structured dataset or database for future training/validation/testing purposes, it will exist in an accurate, high-quality manner that meets standardized requirements.

The CNN architecture is intended for assimilating spatial and temporal hand gesture features through convolution, activation, pooling, and classification processes. Convolutional layers extract appropriate characteristics with activation functions such as ReLU to introduce non-linearity into the convolutional output. Pooling layers help to decrease dimensionality while preserving essential information. Fully connected layers classify features into predefined sign categories. The model is trained with optimization techniques such as Adam and uses evaluation metrics including accuracy, precision, recall, and F1-score to verify the model's performance.

Once trained, CNNs are implemented into a real-time system that receives a continuous stream of live video frames and can be converted into text and additionally, optionally converted into spoken words via the text to speech module. In addition to displaying the text output, the system also displays the associated sign symbol visually to facilitate communication between users. This system also provides multilingual output in three of the major

Indian languages; thereby expanding accessibility to users with various language backgrounds and providing an opportunity for inclusive interaction within the user community.

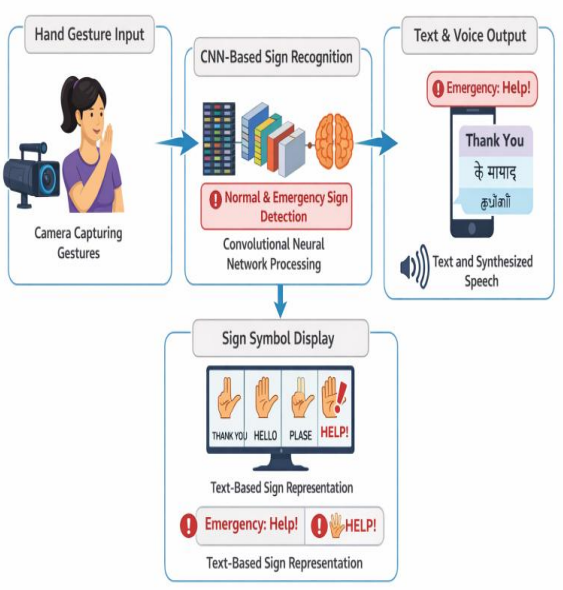


Figure 1: Proposed Architecture

A. Gestures Captured

This module is the starting point for the entire system (i.e., it will capture the hand gestures of the users in real-time). The hand gesture capture module captures the video frames continuously from the camera (either through a webcam or through a mobile device) so that everything is captured without interruption. Once the video frames are captured, each of them becomes a frame of reference. The frames will then be resized to have the same dimensions, and they will also be normalized (i.e., put into the same setting) so that when the gesture recognition systems take the normalized video frames as their input, all the input will be the same size. Noise reduction algorithms (e.g., using filters) are used to reduce any noise caused by the background to enhance the gesture, and other preprocessing techniques are applied to adjust the brightness and enhance the contrast to make sure that gestures are consistently visible regardless of how much light is present in the environment in which they are being performed. Overall, this module provides a high quality consistent (i.e., standard) input to the recognition system. If this module captures the gestures correctly, then the recognition/translating steps will execute properly upon recognizing the gestures, regardless of the environment/user conditions.

B. Gesture Recognition

The gesture recognition module uses a CNN (Convolutional Neural Network) to recognize and classify various hand gestures. Once the hand gestures have been preprocessed, the frames of data are sent to the CNN for processing. The CNN uses multiple convolutional layers to extract spatial features from the frames of data through multiple layers. ReLU activation functions are used on each layer to allow for non-linearity. Pooling layers reduce dimensionality, but still retain significant amounts of information for making effective recognition decisions. Once the CNN has processed the spatial features, fully connected layers classify the hand gestures into predefined categories. The CNN has been trained on a labeled data set that includes a wide variety of hand gesture examples, such as emergency hand signals. During real-time execution of the gesture recognition module, the CNN will analyze live video frames of data on a frame-by-frame basis, with the CNN producing gestures in real-time. This module contains the core intelligence of the overall system; it provides high quality gesture recognition accuracy, high speed gesture recognition processing, and the ability to adapt the gesture recognition system to different users and environments.

C. Text and Voice Conversion

The Gesture Recognition System has the ability to convert recognized gestures into text and voice to facilitate understanding. The gesture identified gets assigned a text and visual representation of the gesture immediately for display to the user. The converted text is then used by a text-to-speech synthesizer to generate understandable spoken output for users that are unable to verbally communicate their message to a person who does not understand sign language. Furthermore, this allows for critical messages, such as an emergency signal, to be communicated clearly and promptly. Also, the low latency processing through system integration creates seamless processing of text-to-speech audio output to enable effective use across different devices by using clear audio settings. This functionality provides the greatest usability/accessibility factor available to both

the end user and the recipient, which is achieved through the melding of both types of outputs.

D. Support for Multiple Languages

Multilingual Support Module offers language translation support for the system capabilities to all users within India regardless of their native language to accommodate possible users having different linguistic regional variations in India by providing translations in these three major Indian languages (Hindi Marathi and Tamil). Once the gesture is recognized by the system, the appropriate text for that gesture will be dynamically assigned to the selected language of the user while ensuring the integrity of the original meaning has not changed with the assignment. The Multilingual Support Module provides all users from different regions throughout India the opportunity to interact with the system and communicate with others. The user can configure the system with the language that the user desires, and then the system will automatically send the output from the text output to the text to speech engine for that selected language. By supporting multiple languages, the user can access the system interface, increase social involvement with other users, and increase the number of ways that the communication process can happen regardless of your native language making the communication as clear and easily understood regardless of the language

E. Visual Display

This module helps to display recognized gestures and outputs in a way that makes each gesture easy to understand, as well as easy to use. The text for each gesture is displayed in a prominent way and often includes the matching sign representation for the user to be able to verify and learn at the same time. In addition, the design of this interface is intended to be clear and simple to use, making it easy for even first-time users to complete their task smoothly. Further, by combining voice output with on-screen visuals of recognized gestures in real time, this interface provides users and listeners a double layer of understanding. The combination of text, symbols, and audio will provide users with increased confidence and ease of access to communicate effectively and intuitively with others who are speech impaired.

## IV. EXPECTED RESULT

The hand gesture-based real-time sign language recognition system identifies and translates gestures into text and sound to support communication for those with speaking difficulties. A gesture recognition model based on convolutional neural networks (CNN) demonstrated a high level of accuracy in gesture classification, successfully recognizing emergency signal gestures and therefore providing reliable real-time gesture identification. A text-to-voice conversion module successfully converted gestures into legible text and intelligible speech that allowed users to communicate with individuals who are unfamiliar with sign language. The multilingual support module successfully converted gesture outputs to text to three primary Indian languages thereby ensuring inclusion of users speaking these languages (ex. Hindi, Tamil, and Urdu). The visual display module enhances users' understanding by showing text display images of sign symbols for each gesture identified and real-time gesture identified. Overall, the system reduced barriers to communication, allowed for fast emergency response by providing visual identification of an emergency or within a social interaction, improved education opportunities, and enhanced access to services. The replies from users in a practical trial indicated an increase in users' confidence, independence, and satisfaction which indicates that the project provides a scalable and accessible solution for bridging communication gaps in everyday and critical situations.

A. Dataset and Input Representation

- In this project, the dataset consists of labeled sign language gesture images or extracted video frames. Each image is represented as a 3D matrix:

$$I \in \mathbb{R}^{H \times W \times C}$$

Where:

- H = Height of the image
- W = Width of the image
- C = Number of channels (1 for grayscale, 3 for RGB)

Before training, pixel normalization is performed:

$$I_{norm} = \frac{I}{255}$$

This scales pixel values between 0 and 1, improving training stability and convergence speed.

B. Convolution Operation

The main operation in CNN is convolution, which extracts spatial features from the image. The convolution output is calculated as:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n) \cdot K(m,n)$$

Where:

- I= Input image
- K = Kernel (filter)
- S(i,j) = Feature map output

This operation helps detect edges, textures, and structural patterns of hand gestures.

C. Activation Function (ReLU)

To introduce non-linearity, the Rectified Linear Unit (ReLU) activation function is applied:

$$f(x) = \max(0, x)$$

ReLU removes negative values and speeds up training by reducing vanishing gradient problems.

D. Softmax Classification

In the final layer, the Softmax function converts outputs into probability values for each sign class:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

Where:

- $z_i$ = Output score for class iii
- C = Total number of gesture classes
- P(yi) = Probability of class iii

The predicted class is determined by:

$$\hat{y} = \arg\max P(y_i)$$

E. Loss Function (Categorical Cross-Entropy)

To measure prediction error during training, cross-entropy loss is used:

$$L = -\sum_{i=1}^{C} y_i \log(P(y_i))$$

Where:

- $y_i$= True label (one-hot encoded)
- P(yi)= Predicted probability

The objective is to minimize this loss using optimization algorithms such as Adam or SGD.

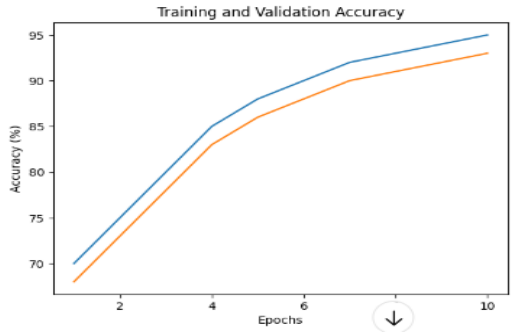


Figure 2; Training and Validation Accuracy

Figure 2 shows the training and validation accuracy of the CNN model over multiple epochs. The graph indicates a steady increase in both accuracies, demonstrating effective learning and good generalization. The small gap between training and validation accuracy suggests minimal overfitting and strong performance of the model.

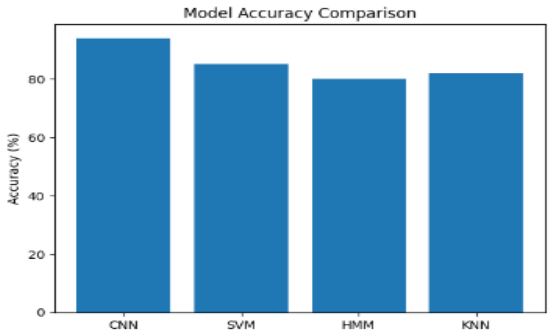


Figure 3 :Model Accuracy

Figure 3 presents the accuracy comparison of different models including CNN, SVM, HMM, and KNN. The CNN model achieves the highest accuracy among all models. This demonstrates that deep learning-based approaches outperform traditional machine learning techniques for sign language recognition tasks.

## V. CONCLUSION

The real-time online sign language recognition system is both practical and efficacious; this is evidenced by its ability to help those who are unable to communicate via spoken language due to their disability, through the identification of hand movements (signs) and translation into text or speech mode, virtually instantly. It relies on a CNN (Convolutional Neural Network) based gesture recognition model, capable of detecting over 200 sign language movements that can be found in Indian Sign Language including signs for emergency situations. The addition of three main Indian languages (Hindi, Tamil, and Telugu) to support multilingual capabilities and the inclusion of visual symbols for the signs further improve accessibility and usability. In addition, by providing users with visual, verbal, and audio outputs, this system allows them to experience increased independence, confidence, and a higher degree of social-interaction. It also fosters greater awareness and understanding of the sign language community for people who may not be part of the sign language community. The project successfully bridges the gap between the needs of those in the deaf and hard of hearing communities and the desire of society to provide communication opportunities for them. By doing this, it has enhanced the capability of those in the deaf and hard of hearing communities to have equitable access to educational, social, and public service opportunities and made it easier for them to communicate with others. Therefore, the rapidly emerging trend toward providing high-quality, scalable, dependable, and user-friendly solutions to assistive technology for people with moderate to severe impairments due to speech and/or communication disorders will provide countless individuals with an opportunity to enhance their quality of life and move towards an inclusive and accessible world.

## REFERENCES

[1] Luqman, Hamzah, et al. "The SignEval 2025 Challenge at the ICCV Multimodal Sign Language Recognition Workshop: Results and Discussion." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.

[2] Toshpulatov, Mukhiddin, et al. "Deep learning pathways for automatic sign language processing." Pattern Recognition 164 (2025): 111475.

[3] Almjally, Abrar, and Wafa Sulaiman Almukadi. "Deep computer vision with artificial intelligence-based sign language recognition to assist hearing and speech-impaired individuals." Scientific Reports 15.1 (2025): 32268.

[4] Ferreira, William, et al. "Automatic Spoken-To-Sign Language Translation: A Review of Key Technologies, Limitations, and Interdisciplinary Opportunities." 2025 27th Symposium on Virtual and Augmented Reality (SVR). IEEE, 2025.

[5] Berrezueta-Guzman, Santiago, Refia Daya, and Stefan Wagner. "Virtual reality in sign language education: opportunities, challenges, and the road ahead." Frontiers in Virtual Reality 6 (2025): 1625910.

[6] Qadhi, Omaimah. "Sign language use in healthcare: professionals' insight." PeerJ 13 (2025): e19446.

[7] Selvaraju, Theetchenya, et al. "Sign language recognition: Enhancing communication for the hearing and speaking impaired using hybrid model." AIP Conference Proceedings. Vol. 3279. No. 1. AIP Publishing LLC, 2025.

[8] Li, Wenwen, et al. "From Crisis to Opportunity: The Barriers to Chinese National Sign Language Acceptance Among Hearing and Speech-Impaired Persons in China Amid the Pandemic." SAGE Open 15.2 (2025): 21582440251332390.

[9] Khan, Asma, et al. "Deep learning approaches for continuous sign language recognition: A comprehensive review." IEEE Access (2025).

[10] Shen, Xin, et al. "Auslanweb: A scalable web-based australian sign language communication system for deaf and hearing individuals." Proceedings of the ACM on Web Conference 2025.

[11] Kılıboz, Nurettin Çağrı, and Uğur Güdükbay. "A hand gesture recognition technique for human–computer interaction." Journal of Visual Communication and Image Representation 28 (2015): 97-104.

[12] Kılıboz, Nurettin Çağrı, and Uğur Güdükbay. "A hand gesture recognition technique for human–computer interaction." Journal of Visual

Communication and Image Representation 28 (2015): 97-104.

[13] Shrivastava, Rajat. "A hidden Markov model based dynamic hand gesture recognition system using OpenCV." 2013 3rd IEEE International Advance Computing Conference (IACC). IEEE, 2013.

[14] Wang, Guojiang. "Facial expression recognition method based on Zernike moments and MCE based HMM." 2016 9th International Symposium on Computational Intelligence and Design (ISCID). Vol. 2. IEEE, 2016.

[15] Bagri, Neelima, and Punit Kumar Johari. "A comparative study on feature extraction using texture and shape for content-based image retrieval." International Journal of Advanced Science and Technology 80.4 (2015): 41-52.

[16] Yun, Liu, Zhang Lifeng, and Zhang Shujun. "A hand gesture recognition method based on multi-feature fusion and template matching." Procedia Engineering 29 (2012): 1678-1684.

[17] Plouffe, Guillaume, and Ana-Maria Cretu. "Static and dynamic hand gesture recognition in depth data using dynamic time warping." IEEE transactions on instrumentation and measurement 65.2 (2015): 305-316.

[18] Chai, Xiujuan, et al. "Two streams recurrent neural networks for large-scale continuous gesture recognition." 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016.

[19] Simão, Miguel A., Pedro Neto, and Olivier Gibaru. "Unsupervised gesture segmentation by motion detection of a real-time data stream." IEEE Transactions on Industrial Informatics 13.2 (2016): 473-481.

[20] Neverova, Natalia, et al. "multi-scale deep learning for gesture detection and localization." European Conference on Computer Vision. Springer, Cham, 2014.