International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

Bridging Silence: ISL Recognition and Speech Generation Using AI

Devi Chithra S[1], Kaaviyaa G[2], Preethiga V[3], Anu J[4]
[1] Assistant Professor, Artificial Intelligence and Data Science, Kamaraj College Of Engineering and Technology, Virudhunagar, Tamil Nadu
[2,3,4]UG- Artificial Intelligence And Data Science, Kamaraj College Of Engineering And Technology, Virudhunagar, Tamil Nadu
***Emails**: Devisankar12@Gmail.Com[1], Ganesankaaviyaa@Gmail.Com[2], Preethigavmdu2021@Gmail.Com[3], Anuswetha2005@Gmail.Com[4]*

Abstract
Indian Sign Language (ISL) is a structured visual-manual language used by more than 18 million deaf and speech-impaired individuals in India. However, limited technological support and accessible digital tools often create communication barriers between deaf and speech-impaired individuals and the general public. This paper presents SaigaiOli, a real-time web-based Indian Sign Language recognition system that translates static hand gestures into text and speech using deep learning and computer vision techniques. The proposed system recognizes 36 ISL gesture classes, including 26 alphabets (A-Z) and 10 numerical signs (0-9). A Kaggle-sourced dataset containing approximately 2,000 images per class was used for training, resulting in a total of about 72,000 images collected from signers of different age groups. Hand landmarks are extracted using the Media-Pipe Hands framework, producing 84 normalized features per gesture. These features are classified using a sequential deep neural network implemented with Tensor-Flow. The model achieved an overall accuracy of 96.75%, with macro-averaged precision, recall, and F1-score of 0.97. The system is deployed as a React.js web application, enabling real-time gesture recognition through a webcam, text accumulation, text-to-speech output, and an integrated ISL learning module. The proposed system promotes accessible communication and supports greater inclusion for deaf and speech-impaired individuals.

***Keywords:** Deep learning; Hand gesture recognition; Indian Sign Language; Media-Pipe; Real-time recognition*

1. Introduction
Sign languages are natural, fully expressive languages used by Deaf and speech impaired communities worldwide. Indian Sign Language (ISL) is the primary language of the Indian Deaf and speech-impaired community, estimated at over 18 million users across India. Despite this widespread use, ISL lacks adequate digital infrastructure, creating significant barriers to communication, education, and social inclusion for its users. Automated Sign Language Recognition (SLR) has attracted substantial research attention due to advances in deep learning and computer vision. Early Continuous Sign Language Recognition (CSLR) approaches focused on RGB video and optical flow but struggled with fine-grained spatiotemporal feature extraction across frames. (Z. Wang et al., 2025) addressed this with STNet, a spatial-temporal feature-enhanced network using an optimal transport-based spatial resonance module and a multi-temporal perception decoder, achieving a 2.9% improvement over state-of-the-art methods on PHOENIX14 and CSL-Daily benchmarks [1]. Similarly, (Hu et al., 2023) proposed STFE-Net, which fuses spatial pose features from 53 abbreviated key-points with a Transformer-based temporal encoder, achieving BLEU-4 scores of 72.14 and 22.45 on custom and PHOENIX14-T datasets [8]. For isolated static gesture classification, lightweight CNN architectures have shown strong performance. (Reddy et al., 2025) proposed a real-time system using MobileNetV1 with transfer learning on a 21-class dataset, achieving 97.14% test accuracy with weighted F1-score of 0.97, demonstrating the effectiveness of lightweight architectures for real-time deployment [2]. (Dabwan

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

et al., 2024) used MobileNetV2 for ASL alphabet recognition across 24 classes, achieving 99.9% accuracy on 27,455 training samples [5]. Skeleton and pose-based approaches offer robustness against lighting and background variation. (Miah et al., 2023) proposed GCAR, a two-stream multistage graph convolution network with attention and residual connections on joint skeleton and motion streams, achieving 90.31% and 99.75% accuracy on WLASL and MSL datasets respectively [6]. (Naz et al., 2023) developed Sign-Graph, a lightweight GCN pipeline using hands and body pose, outperforming prior pose-based methods by up to 27.62% on WLASL subsets and achieving 100% on LSA-64 [7]. Media-Pipe, developed by Google, provides robust real-time hand landmark detection extracting 21 3D key-points per hand from a standard RGB camera. (Yoga et al., 2024) combined CNN, RNN, and Media-Pipe hand detection for real-time sign interpretation from webcam video, demonstrating low-latency processing and strong adaptability to diverse lighting conditions [3]. (Ihsan et al., 2024) proposed MediSign, an attention-based CNN-BiLSTM system using MobileNetV2 for medical word-level sign recognition between doctors and patients, achieving 95.83% validation accuracy across 30 words from 20 diverse signers [4]. Motivated by these advances, this paper presents SaigaiOli - a browser-deployable, real-time ISL recognition web application. The primary contributions are: (1) an end-to-end ISL recognition pipeline deployable on standard web browsers without specialized hardware, (2) a training dataset of ~72,000 images across 36 ISL classes with diverse age group representation achieving 96.75% overall accuracy, and (3) an accessible interactive learning platform designed to celebrate and empower the ISL community.

1.1. Indian Sign Language

ISL is a visual-gestural language with distinct grammar and syntax, independent of spoken Indian languages. This work focuses on static hand-shape recognition for 26 alphabets (A-Z) and 10 digits (0-9), forming the foundational vocabulary for fingerspelling-based digital communication.

1.2. Motivation And Scope

Most existing ISL recognition systems are research prototypes requiring specialized hardware or desktop installation. SaigaiOli addresses this by delivering recognition through a standard web browser with only a webcam, serving as both a real-time communication aid and an interactive ISL learning tool.

2. Method

The proposed system consists of three main components: (1) landmark-based feature extraction using Media-Pipe Hands, (2) a deep neural network classifier, and (3) a full-stack web application for real-time deployment. The overall system design is illustrated in Figure 1 and the technical architecture in Figure 2.

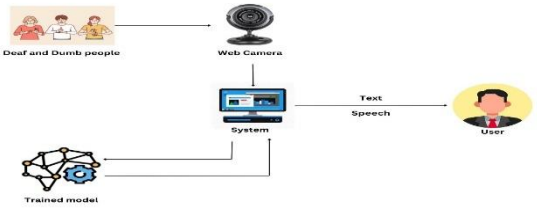


Figure 1. System Design Of Saigaiol - Deaf And Speech-Impaired Users Perform ISL Signs In Front Of A Webcam; The System Processes The Feed Through A Trained Deep Learning Model And Delivers Text And Speech Output To The User.

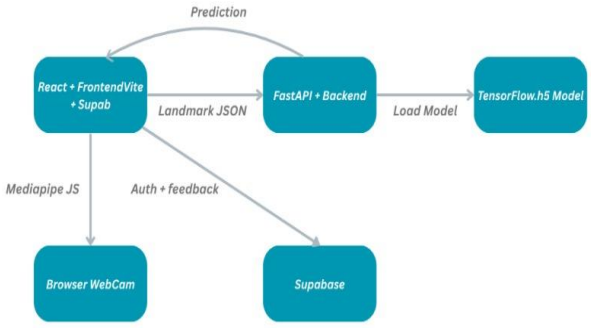


Figure 2 Technical Architecture Of Saigaioli - The React.Js Frontend Communicates With The Fastapi Backend Via Landmark Json And Receives Predictions, While Supabase Handles Authentication And Feedback, And Media-Pipe Js Runs Client-Side In The Browser.

2.1.Dataset

The dataset was sourced from Kaggle, comprising approximately 2,000 images per class across all 36

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

ISL sign categories, resulting in a total of approximately 72,000 images. The dataset includes gestures performed by children, adults, and elderly individuals of varying hand sizes and skin tones, ensuring demographic diversity for robust real-world generalization - a critical factor highlighted by (Dabwan et al., 2024) [5]. The dataset was split with 80% for training and 20% for testing, consistent with standard evaluation protocols shown in Table 1.

Table 1 Dataset Summary Per Class Category

Category	Classes images per class	Total images per class	Features Dimensions
Alphabets	26 (A-Z)	~2,000 ~52,000	84
Digits	10 (0-9)	~2,000 ~20,000	84
Total	36	~2,000 ~72,000	84

2.2.Feature Extraction

Each image is processed using Media-Pipe Hands in Python. For each detected hand, 21 landmarks with x and y coordinates are extracted, producing 42 features per hand. For two-hand support, both hands are concatenated to produce an 84-dimensional feature vector. Coordinates are normalized relative to the minimum x and y values within each hand to ensure scale and position invariance across different hand sizes and camera distances - consistent with skeleton-based methods of (Miah et al., 2023; Naz et al., 2023) [6&7]. The extracted feature arrays and labels are serialized into a pickle file for model training.

2.3.Model Architecture

A sequential deep neural network is implemented using Tensor-Flow/Keras, consisting of fully connected Dense layers with ReLU activations and Dropout regularization. The output layer uses Soft-max activation for 36-class classification, trained using the Adam optimizer with categorical cross-entropy loss. This dense, landmark-based architecture is deliberately lightweight compared to graph convolution networks of (Hu et al., 2023; Wang et al., 2025) [1&8], prioritizing inference speed for real-time browser deployment. The trained model is exported as an .h5 file and a Label-Encoder is serialized via pickle for class label mapping.

2.4.Web Application

The SaigaiOli web application is built with React.js and Vite, deployed on Vercel at saigaioli.vercel.app. Media-Pipe Hands runs entirely client-side using the @mediapipe/hands JavaScript package, enabling real-time landmark extraction without server video streaming - consistent with the low-latency principles of (Yoga et al., 2024) [3]. Extracted 84-dimensional feature vectors are sent to a Fast-API backend via HTTP POST for model inference, throttled to one prediction per 800 milliseconds. Supabase provides email/password authentication and feedback database storage. The application comprises four modules: (1) Home - introduction and platform mission, (2) Live Capture - real-time gesture recognition with landmark overlay, text accumulation, and Auto-Speak, (3) Learn ISL - reference image cards for all 36 signs with click-to-speak, and (4) Feedback - user experience rating and suggestions.

3. Results And Discussion

3.1 Results

3.1.1 Model Evaluation

The trained deep neural network was evaluated on the 20% held-out test set comprising 67,760 samples across all 36 classes. The model achieved an overall accuracy of 96.75%. The macro-average and weighted-average precision, recall, and F1-score all reached 0.97, demonstrating strong and consistent performance across all sign categories shown in Table 2.

Table 2 Model Evaluation Summary

Metric	Score
Overall Accuracy	96.75%
Macro Average Precision	0.97

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

Macro Average Recall	0.97
Macro Average F1-Score	0.97
Weighted Average Precision	0.97
Weighted Average Recall	0.97
Weighted Average F1-Score	0.97

Table 3 Per-Class Classification Report (Selected Signs)

Sign	Precision	Recall	F1-Score	Support
0	0.96	0.88	0.92	2022
1	0.97	0.95	0.95	1889
5	0.99	0.97	0.98	1840
7	0.99	0.99	0.99	1968
9	0.98	0.99	0.99	3811
A	1.00	1.00	1.00	1696
B	0.99	0.99	0.99	1803
C	0.99	0.99	0.99	1642
E	0.98	0.97	0.97	3142
M	0.88	0.87	0.87	1695
N	0.90	0.88	0.88	1797
P	0.99	0.99	0.99	2000
W	0.97	0.98	0.98	1743
Z	1.00	0.99	0.99	2290

The highest-performing signs are A (F1: 1.00), C (F1: 0.99), Z (F1: 0.99), and digits 7, 8, 9 - all achieving F1-scores at or above 0.99. The most challenging signs are M (F1: 0.87) and N (F1: 0.88), which are visually similar static hand-shapes with overlapping finger configurations, a known difficulty in both ISL and ASL recognition literature shown in Table 3.

3.1.2 Application Interface

The SaigaiOli web application interface consists of four pages. Figure 3 shows the Home page, which introduces the platform's purpose of empowering the Indian Deaf and speech-impaired community. Figure 4 shows the Live Capture page prior to camera activation, displaying the two-panel layout with the camera feed area and the prediction/text accumulation panel. Figure 5 shows the Learn ISL page with reference images for all 36 signs categorized by All Signs, Alphabets (A-Z), and Numbers (0-9), with click-to-speak functionality shown in Figure 3, 4 and 5.

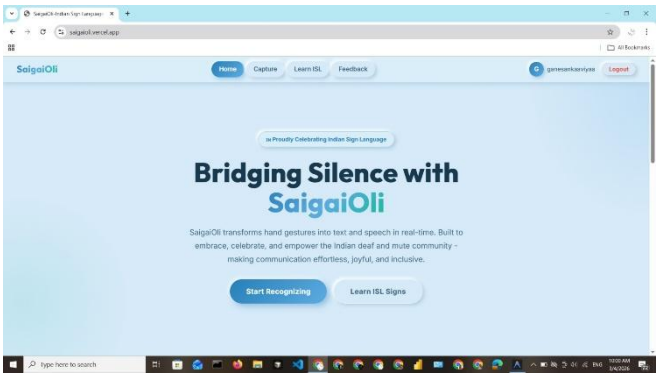


Figure 3 SaigaiOli Home Page - Introducing The Platform Mission Of "Bridging Silence" For The Indian Deaf And Speech-Impaired Community, With Navigation To Capture And Learn ISL Modules.

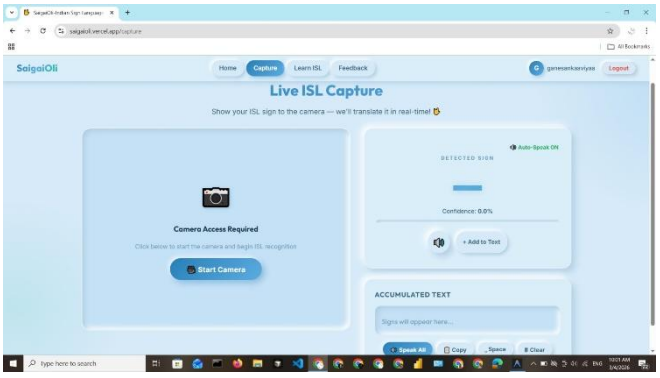


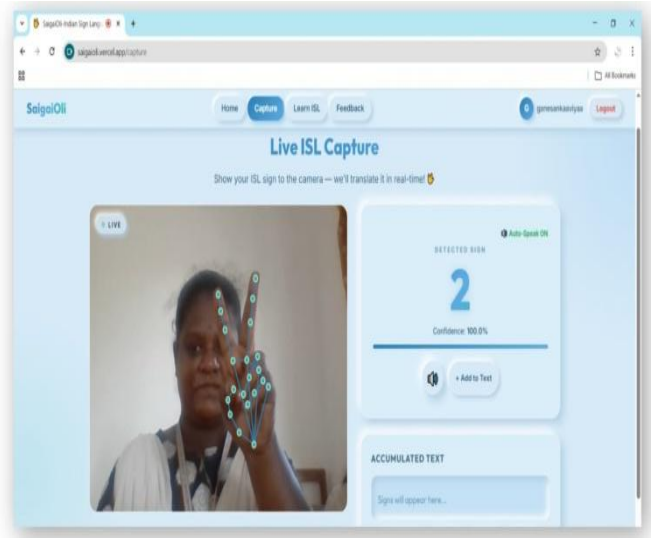
Figure 4 SaigaiOli Live ISL Capture Page - Two Panel Layout Showing The Camera Access Prompt With Start Camera Button On The Left And The Detected Sign Panel With Auto-Speak, Add To Text, And Accumulated Text Controls On The Right.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

Figure 5 Saigaioli Learn ISL Page - Reference Image Cards For All 36 ISL Signs (Digits 0-9 And Alphabets A-Z) With Filter Tabs And Click-To-Speak Functionality, Displaying Demographically Diverse Signers Including Children, Adults, And Elderly Individuals.

3.1.3 Live Recognition Results

Figures 6, 7 and 8 demonstrate real-time recognition during live deployment. Figure 6 shows the signs '2' and 'W' recognized with 100% confidence from adult users, with the Media-Pipe hand landmark skeleton overlaid. Figure 7 presents sequential recognition of the characters forming "OK" (O → K) and the digits "17" (1 → 7), demonstrating the system's capability to correctly interpret both alphabetic and numeric gestures in real time. Figure 8 illustrates the text accumulation feature spelling "ZEBRA" sequentially via Z → E → B → R → A shown in Figure 6,7 and 8.



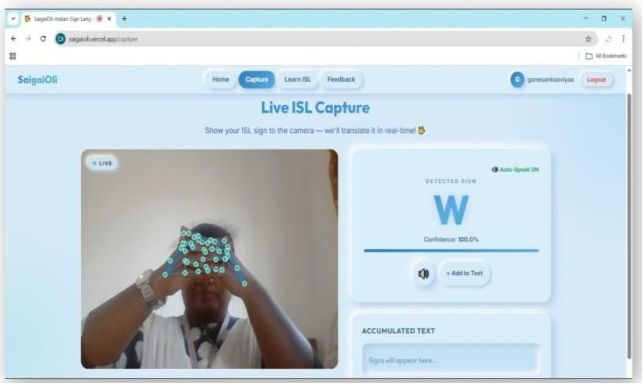


Figure 6 Real-time ISL recognition - Media-Pipe hand landmarks overlaid on webcam feed for signs '2' and 'W', both detected at 100% confidence, demonstrating accurate adult user recognition.

Figure 7 Sequential Recognition Feature Detecting O, K, 1, and 7 To Form "Ok" And "17"

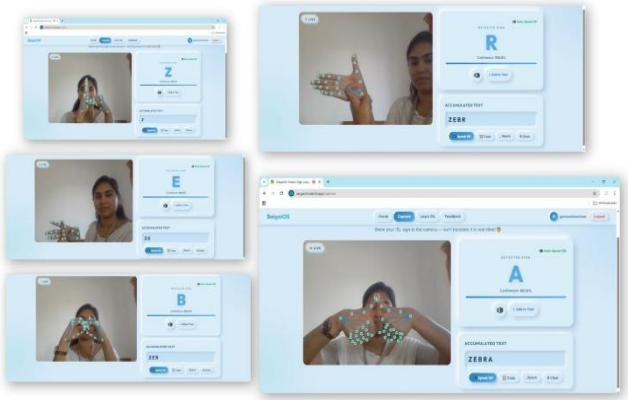


Figure 8 Live Text Accumulation Feature Sequentially Recognizing Z, E, B, R, A To Form "Zebra"

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

3.2 Discussion

The overall accuracy of 96.75% with macro F1-score of 0.97 validates the effectiveness of the Media-Pipe landmark-based feature representation for ISL static hand-shape classification. The normalized relative coordinate approach generalizes well across the demographically diverse dataset - children, adults, and elderly signers, as evidenced by live deployment results showing consistent 100% confidence scores for well-formed gestures.

Compared to the pixel-based MobileNetV1 approach of (Reddy et al., 2025) (97.14% on 21 classes) and MobileNetV2 approach of (Dabwan et al., 2024) (99.9% on 24 classes), SaigaiOli achieves competitive accuracy (96.75%) on a more challenging 36-class dataset while eliminating the need for image preprocessing, GPU inference, or server-side video processing [2][5]. The client-side Media-Pipe execution delivers low-latency real-time performance aligned with the accessibility objectives of (Yoga et al., 2024) [3]. The most challenging signs - M (F1: 0.87) and N (F1: 0.88) - share similar hand-shapes where fingers fold over the thumb in slightly different configurations, a known ambiguity in static sign recognition. This is consistent with observations in (Miah et al., 2023), where finger-level spatial ambiguity was identified as a primary challenge for landmark-based methods without temporal context [6]. The confusion matrix further confirms that most misclassifications occur between visually similar signs (M/N, O/C) rather than across distinct sign categories, indicating the model has learned meaningful ISL hand-shape representations. The Learn ISL module serves a dual purpose, it provides an educational reference for learners and implicitly validates the dataset quality, as reference images from the Kaggle dataset showing diverse age groups (children, adults, elderly) are displayed for all 36 supported signs. This demographic diversity in both the training dataset and the reference module is a distinguishing feature of SaigaiOli compared to most existing SLR research prototypes, which typically train on a single demographic group. Future work will extend the system to dynamic ISL word-level recognition using temporal models, potentially incorporating spatiotemporal fusion from STNet (Wang et al., 2025) or the Transformer-based temporal encoding from STFE-Net (Hu et al., 2023) [1][8] and will include additional ISL vocabulary beyond static fingerspelling.

Conclusion

This paper presented SaigaiOli, a browser-deployable real-time Indian Sign Language recognition web application supporting 36 ISL signs (A-Z, 0-9), publicly accessible at saigaioli.vercel.app. The system integrates Media-Pipe based hand landmark extraction, a trained Tensor-Flow deep neural network served via Fast-API, and a React.js frontend with text-to-speech output deployed on Vercel. The model achieves an overall accuracy of 96.75% with macro-average precision, recall, and F1-score of 0.97 on a demographically diverse Kaggle dataset of approximately 72,000 images spanning children, adult, and elderly signers. Live deployment testing confirms consistent recognition with confidence scores reaching 100% for well-formed gestures across multiple real-world users. By combining real-time recognition, sequential text accumulation, Auto-Speak voice output, and an interactive ISL learning module, SaigaiOli bridges the communication gap for the Indian Deaf and speech-impaired community, promoting digital accessibility and inclusivity without requiring any specialized hardware. Future enhancements will target dynamic word-level ISL gesture recognition by incorporating spatiotemporal deep learning models.

Acknowledgements

The authors would like to thank the Department of Artificial Intelligence and Data Science, Kamaraj College of Engineering and Technology, for providing the necessary facilities and support to carry out this research work. The authors also express their sincere gratitude to all the faculty members and colleagues who provided valuable suggestions during the development of this study. The authors would like to thank the editors of IMSTEM and the anonymous reviewers for their time and valuable review of this manuscript.

References

[1]. Wang, Z., Li, D., Jiang, R., & Okumura, M. (2025). Continuous sign language recognition with multi-scale spatial–temporal feature enhancement. IEEE

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

Access, 13, 5491–5506. doi: 10.1109/ACCESS.2025.3526330.

[2]. Reddy, V. H., Kovela, N. S., N., N., & Agrawal, S. (2025). Sign language recognition using MobileNetV1: A real time approach. Proceedings of the 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), 1–6. doi: 10.1109/IEMENTech65115.2025.10959470.

[3]. Yoga, M., Ramyasri, M. M., Raj, B. H., Gokul, G., Praveen, E. K. R., & Angamuthu, S. (2024). Sign language detection using deep learning. Proceedings of the 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), 751–756. doi: 10.1109/ICDICI62993.2024.10810892.

[4]. Ihsan, M. A., Eram, A. F., Nahar, L., & Kadir, M. A. (2024). MediSign: An attention-based CNN-BiLSTM approach of classifying word level signs for patient–doctor interaction in hearing impaired community. IEEE Access, 12, 33803–33815. doi: 10.1109/ACCESS.2024.3370684.

[5]. Miah, A. S. M., Hasan, M. A. M., Nishimura, S., & Shin, J. (2024). Sign language recognition using graph and general deep neural network based on large scale dataset. IEEE Access, 12, 34553–34569. doi: 10.1109/ACCESS.2024.3372425.

[6]. Dabwan, B. A., Jadhav, M. E., Yami, M. A., Hassan, E. A., Almula, S. M., & Ali, Y. A. (2024). Classifying hand gestures for people with disabilities utilizing the MobileNetV2 model. Proceedings of the 1st International Conference on Innovative Sustainable Technologies for Energy, Mechatronics, and Smart Systems (ISTEMS), 1–4. doi: 10.1109/ISTEMS60181.2024.10560333.

[7]. Hu, J., Liu, Y., Lam, K. M., & Lou, P. (2023). STFE-Net: A spatial-temporal feature extraction network for continuous sign language translation. IEEE Access, 11, 46204–46217. doi: 10.1109/ACCESS.2023.3234743.

[8]. Naz, N., Sajid, H., Ali, S., Hasan, O., & Ehsan, M. K. (2023). SignGraph: An efficient and accurate pose-based graph convolution approach toward sign language recognition. IEEE Access, 11, 19135–19147. doi: 10.1109/ACCESS.2023.3247761.

[9]. Poladiya, P., Suresh, D., Gulhane, P., Ajmal, M. A., & Kosamkar, P. (2024). Sign language detection using deep learning. Proceedings of the 3rd International Conference for Innovation in Technology (INOCON), 1-6. doi:10.1109/INOCON60754.2024.10512307

[10]. Mohebbanaaz, A., Babu, R., Himaja, Y. S., Stuthi, S. J., & Mamatha, B. (2024). Transformation of sign language to text in digital era using deep neural network. Proceedings of the IEEE International Conference on Computational Intelligence and Communication Networks (CICN), 1278-1283. doi:10.1109/CICN63059.2024.10847557.

[11]. Deshmukh, R., Lahamge, T., Phadatare, I., Shinde, D., & Manhas, R. (2024). Real-time Marathi sign language recognition using deep learning techniques. Proceedings of the International Conference on Automation, Computing and Renewable Systems (ICACRS), 1326-1331. doi: 10.1109/ICACRS62842.2024.10841608.

[12]. A.J. S., R. R., J. B., D. V. K., L. G. P., & S. H. (2024). Real-time American Sign Language gesture recognition using transfer learning. Proceedings of the International Conference on Circuit Power and Computing Technologies (ICCPCT), 1604-1609.doi:10.1109/Iccpct61902.2024.10672721.

[13]. Devabathini, N., & Mathivanan, P. (2023). Sign language recognition through video frame feature extraction using transfer learning and neural networks. Proceedings of the International Conference on Next Generation Electronics (NEleX), 1-6. doi:10.1109/NEleX59773.2023.10421383.

[14]. S., E., N., A., S., H. S., & J., H. (2023). Hand gesture vocalizer using MobileNetV2 for deaf mutes. Proceedings of the International Conference on Pervasive Computing and Social Networking (ICPCSN), 1481-1485. doi:10.1109/ICPCSN58827.2023.00246.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2004-2011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0267

[15]. Singla, N. (2023). American Sign Language letter recognition from images using CNN. Proceedings of the International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 1-9. doi: 10.1109/ICEEICT56924.2023.10156922.

[16]. Sayeed, M. A., Islam, M. S., Islam, M. B., Pareek, P. K., & Rohan, T. I. (2023). Traffic sign recognition and classification using CNN with transfer learning. Proceedings of the International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), 1-5. doi: 10.1109/ICDCECE57866.2023.10151254.

[17]. Gupta, U., Sharma, S., Jyani, U., Bhardwaj, A., & Sharma, M. (2022). Sign language detection for deaf and dumb students using deep learning: Dore idioma. Proceedings of the International Conference on Innovative Sustainable Computational Technologies (CISCT), 1-5. doi:10.1109/CISCT55310.2022.10046657.

[18]. Baumgartl, H., Sauter, D., Schenk, C., Atik, C., & Buettner, R. (2021). Vision-based hand gesture recognition for human-computer interaction using MobileNetV2. Proceedings of the IEEE Annual Computers, Software, and Applications Conference (COMPSAC), 1667-1674. doi:10.1109/COMPSAC51774.2021.00249.

[19]. Setyono, N. F. P., & Rakun, E. (2019). Recognizing word gesture in sign system for Indonesian language sentences using DeepCNN and BiLSTM. Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS), 199-204. doi: 10.1109/ICACSIS47736.2019.8979772.