ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

SIGN-TO-SPEAK: A VISION-BASED REAL-TIME SIGN LANGUAGE TO TEXT AND SPEECH TRANSLATION SYSTEM

Dr. Rasika Kulkarni[1], Gulafsha Sayyed[2], Sanmeetkaur Gulati[3], Pooja Najardhane[4]
[1,2,3,4] *Department of Computer Science, Fergusson College (Autonomous), Pune, India.*
Email: rasika.kulkarni@fergusson.edu[1], gulafshasayyed85@gmail.com[2], anmeetgulati41121@gmail.com[3], najardhanepooja0@gmail.com[4]

Abstract
Sign language serves as the principal mode of communication for millions of deaf and hard-of-hearing individuals globally. The linguistic divide between signers and non-signers presents considerable obstacles in everyday communication. This study introduces Sign-to-Speak, a real-time vision-based system that identifies hand gestures via Mediapipe landmark extraction, uses a Convolutional Neural Network (CNN) to classify them and converts them into the appropriate text and speech. The system effectively manages 26 ASL fingerspelling gestures with significant resilience across varying background and varying lightning conditions. In order to resolve visually similar signs, a significant contribution is the incorporation of skeleton-based gesture normalization with CNN-based classification and rule-based disambiguation. In controlled settings, the model achieves up to 99% accuracy; in real-world situations, it achieves 97% accuracy. The methodology, dataset construction, gesture classification pipeline, results, and limitations are described in this research paper, along with future prospects for continuous sentence formation and dynamic gesture recognition.

Keywords: Sign Language Recognition (SLR), American Sign Language (ASL), Hand Gesture Processing, Convolutional Neural Networks (CNN), Mediapipe, Real-Time Translation, Speech Synthesis, Human-Computer Interaction (HCI).

► *Corresponding Author: Dr. Rasika Kulkarni*

Motivation

Globally there is a population of over 70 million that use sign languages as their primary means of communication. While sign languages convey a wide range of expression, they are also foreign to the great majority of hearing people which in turn creates large scale communication issues in the areas of education, health care, public services and day to day life. To date we have traditional SLR systems which require expensive hardware like data gloves or depth sensors which in turn limits access. The motivation behind this work is to put forth a low cost, real time, camera-based system which will be able to recognize hand gestures, convert them into text, and produce voice output. By which we focus to enable better interaction between the deaf/mute community and non-signers which in turn will promote social inclusion, accessibility and independence.

1. Introduction

In the intersection of deep learning, computer vision and sign language recognition is becoming a focal point of study [1,8,15]. The American Sign Language (ASL) finger spells letters using different hand shapes and motions, and expresses words and phrases using different configurations

of the hands [2,3,5]. With that said, automated recognition is currently in progress [6,7,12]. Key challenges include: varying conditions of light and background, different shapes and skin tones among signers, and the inherent ambiguity of certain gestures [8,9,10,11]. As a result, real-world deployment of Sign Language Recognition systems is still unsatisfactory [5,7,12]. Most of the existing approaches rely on background-dependent features such as contours and binary masks [1,5,6]. The current study aims to fill the gap using the Mediapipe SLR framework to perform 21-point hand landmark tracking to produce personalized skeletons that are background invariant. To classify the subsets of visually similar groups (e.g., M/N/S/T) of alphabets, a combination of hand-made geometric rules and a CNN classifier is used [4,10,14].

2. Literature Review

To capture the latest advancements and shortcomings of SLR systems, 15 research papers were reviewed and gaps were also identified.

Convolutional neural networks (CNNs) were used in early deep learning attempts for sign language recognition (SLR) to extract spatial features under controlled conditions. Pigou et al. (2014) validated deep learning for gesture recognition by confirming CNNs' capacity to capture performance variations across sign video datasets. However, limitations included lab-confined evaluations, small signer numbers, restricted lexicons, and no explicit temporal modelling or real-world testing [1]. Huang et al. (2015) extended this to 3D CNNs for spatio-temporal feature extraction, outperforming 2D methods in temporal dynamics but facing high computational costs, curated video requirements, and gaps in signer independence and environmental robustness [2].

Integration of recurrent neural networks (RNNs) addressed temporal coherence for continuous signing. Cui et al. (2017) introduced recurrent CNNs with staged optimization, enhancing performance over isolated models, though benchmark-limited and sensitive to inter-signer variation [3]. Shah (2018) proposed an end-to-end CNN-encoder-decoder + LSTM pipeline, achieving high accuracy on isolated signs, yet signer-dependent with untested scalability to full systems [4]. Real-time systems emerged, as in Taskiran et al. (2018), demonstrating low-latency edge deployment for restricted ASL sets, but confined to lab settings with limited signers and no diverse viewpoint generalization [5].

By 2020, applications expanded to translation and real-time recognition. Abiyev et al. (2020) developed CNNs for fingerspelling-to-text on pose-varied datasets, showing translation potential, though restricted to static/short sequences without full sentence-level or continuous handling[6]. Kadhim & Khamees (2020) achieved strong accuracy on real ASL datasets via CNNs in controlled conditions, but performance remained untested amid occlusions, lighting variations, or camera angles [7].

Comprehensive reviews highlighted trends and deficits. Wadhawan & Kumar (2021) surveyed a decade of SLR, identifying lacks in standardized benchmarks, signer-independent models, continuous solutions, and user-cantered frameworks [8].

Recent hybrid and lightweight approaches built on these foundations. Sundar & Bagyammal (2022) combined MediaPipe landmarks with LSTM for ~99% ASL alphabet accuracy, proving skeleton-based efficacy, but limited to controlled single-hand letters without signer variation analysis [9]. Jayanthi et al. (2023) fused keyframe extraction, CNN, and LSTM for continuous gestures, improving temporal consistency and efficiency, yet untested on large vocabularies or spontaneous signing [10]. Vyavahare et al. (2023) applied LSTMs to Indian Sign Language (ISL) dynamic gestures on custom datasets, though small-scale with unresolved signer independence and regional variants [11]. Alsharif et al. (2023) and Triwijoyo et al. (2023) deployed modern

CNN/RNNs for ASL alphabets and gesture classification, achieving high accuracy in lab constraints but without word-level or real-world scalability [12,13]. Paul et al. (2024) optimized CNN-LSTM with Adam for real-time ASL (mid-90s% accuracy), but overlooked large-scale continuous evaluation and Deaf user accessibility [14].

Tao et al. (2024) reviewed traditional-to-modern SLR, noting trends toward multimodal fusion and transformers, while pinpointing gaps in data sparsity, linguistic awareness, signer independence, multimodal integration, and resource-constrained deployment [15].

Persistent challenges across include controlled lab data, small vocabularies/signers, signer dependence, inadequate continuous/linguistic modelling, and real-world robustness deficits (e.g., occlusion, lighting, viewpoints), motivating advanced, unconstrained SLR systems.

3. Methodology

A multi-stage pipeline combining computer vision, deep learning, and human-computer interaction components is used to construct the suggested system. The methodology consists of:

3.1 System Flowchart

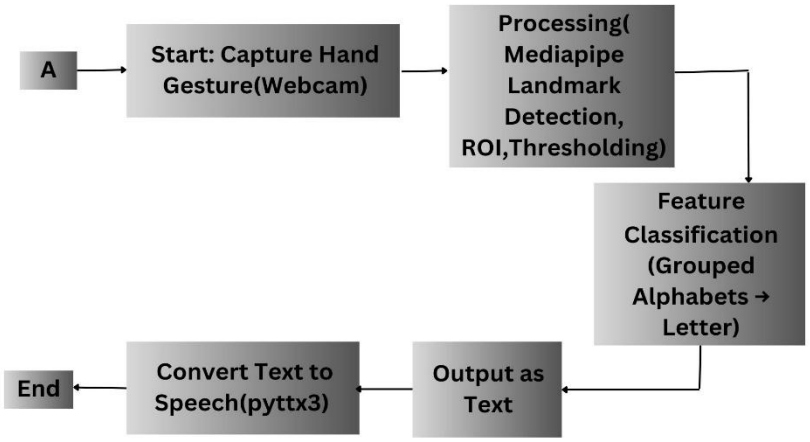


Fig.1.Flowchart

3.2 Hand Landmark Visualization

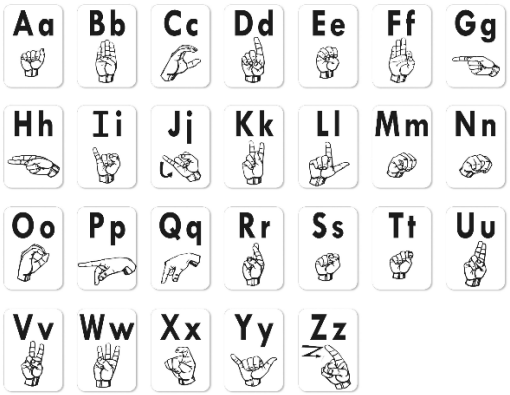


Fig.2.American Sign Language

3.3 Dataset Description

To train and evaluate the proposed Sign-to-Speak system, a **custom dataset** was constructed using real-time webcam recordings of American Sign Language (ASL) static alphabet gestures. The dataset constains **26 distinct classes**, representing the letters A through Z. For each alphabet, approximately **180 samples** were captured, which totals nearly 4,700 images. Unlike conventional SLR datasets that rely on raw RGB frames or background-dependent segmented images [1,2,3,4], this dataset was designed to be **background-invariant**. Using the Mediapipe Hand Landmark model [9], each frame was processed to obtain 21 key hand landmarks. These landmarks were plotted on a standard 400 x 400pixel white canvas to create a hand drawing of the gesture. The approach removes the impact of environmental factors, such as background clutter, skin colour variation, shadows, and changes in lighting. Therefore, the dataset achieved a uniformity and robustness which greatly improved model generalization and performance in real-time.

3.4 Preprocessing

The preprocessing pipeline was designed to convert the raw webcam frames into shaped, structure, and light invariant representations suitable for deep learning [1,2,8,15]. Each frame obtained was first converted into a NumPy array to allow easy manipulation and compatibility with computer vision operations.The system utilized Mediapipe's palm detection and hand landmark extraction modules to equally and accurately identify the hand section and bounding box in varying lighting conditions. After a hand was identified, the system recorded 21 unique positions that capture the locations of all the important joints, like the fingertips and knuckles [9]. To provide a rough appearance of the hand gesture, the system used these position coordinates to construct a skeletal figure [9,10]. On a blank white canvas, the system used a set of straight lines to draw the selected positions and form the structure of the hand, ignoring the noise of the background. This process efficiently removes any background information that is irrelevant to the gesture [1,5,6]. To make sure all hand drawings were from the same reference point in the centre of the canvas we applied positional offsets and normalized the sketches which in turn did not matter the size of the hand or the distance to the camera [3,11,12].

Then we resized and transformed the final images to fit the input requirements of the CNN model [4,10,13,14]. This multi-step preprocessing we did guarantees that the model uses clean and standardized inputs which at the same time preserve gesture integrity and is not affected by lighting, orientation, and background variation [8,9,15].

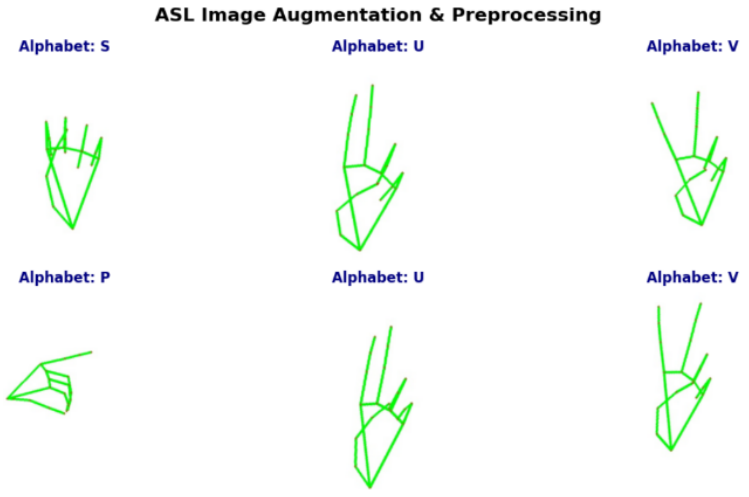
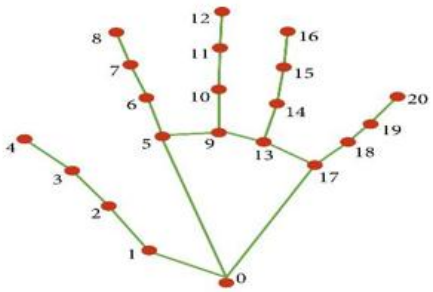


Fig.3.ASL Image



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Fig.4.Landmark of the hand

3.5 Model Building

The classification component of the Sign-to-Speak system is powered by a custom-designed **Convolutional Neural Network (CNN)** optimized for processing skeletal gesture representations [1,4,12,13,14]. The network begins with a stack of convolutional layers that identify simple features like edges, and lines, and the orientation of joints [1,12,13]. After each convolutional layer, a ReLU activation function gets added to ensure non-linearity to the network so that it can learn more complicated functions. Further along the pipeline, the layers that perform max pooling decrease the spatial dimensions within the feature maps, but they do so while retaining the most important features to improve compute efficiency and lessen the chance of overfitting [1,4,10].

To improve model generalization, we see to it that dropout layers are used which at train time randomly drop out a fraction of the neurons' activation which in turn prevents co adaptation of neurons [4,10,14]. After we get high level features out of the network which we do by passing the input through the network layers that extract these features out of the image, we flatten the feature maps out and pass them on to at least one fully connected (dense) layer [12,13]. The final layer we use is the softmax which, in this case, is used to classify the sign into one of 8 large scale gesture categories that were defined during our problem simplification phase [5,6,12,14]. We have designed a model that does a great job at providing the right amount of expressiveness, computational speed and robust design that in turn allows the model to perform in real time on that which is to say off of the average consumer hardware [5,7,14,15].

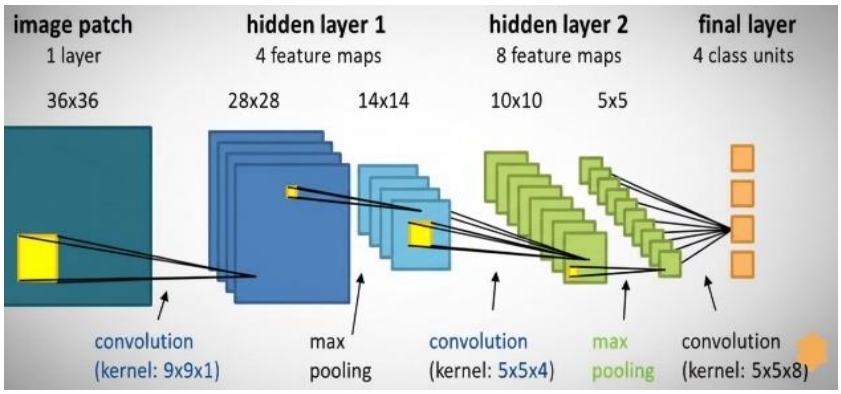


Fig.5.CNN Architecture

Model: "sequential_3"

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 398, 398, 32)	896
max_pooling2d_12 (MaxPooling2D)	(None, 199, 199, 32)	0
conv2d_13 (Conv2D)	(None, 197, 197, 32)	9,248
max_pooling2d_13 (MaxPooling2D)	(None, 98, 98, 32)	0
conv2d_14 (Conv2D)	(None, 96, 96, 16)	4,624
max_pooling2d_14 (MaxPooling2D)	(None, 48, 48, 16)	0
conv2d_15 (Conv2D)	(None, 46, 46, 16)	2,320
max_pooling2d_15 (MaxPooling2D)	(None, 23, 23, 16)	0
flatten_3 (Flatten)	(None, 8464)	0
dense_12 (Dense)	(None, 128)	1,083,520
dropout_6 (Dropout)	(None, 128)	0
dense_13 (Dense)	(None, 96)	12,384
dropout_7 (Dropout)	(None, 96)	0
dense_14 (Dense)	(None, 64)	6,208
dense_15 (Dense)	(None, 8)	520

Total params: 1,119,720 (4.27 MB)
Trainable params: 1,119,720 (4.27 MB)
Non-trainable params: 0 (0.00 B)

Fig.6.CNN Model Summary

4. Results

The findings show that the combination of skeleton-based preprocessing and CNN classification provides reliable and robust gesture recognition.

4.1 Accuracy

The Sign-to-Speak system achieves strong performance across both controlled and real-world environments. In controlled conditions, the CNN classifier reaches **99% accuracy**, demonstrating excellent learning of hand-shape patterns. In real world use which includes various lighting conditions and background activities the system reports 97% accuracy which is very good. That small performance difference which we see is a proof of the skeleton-based representation which we used which in turn minimized the lighting and background noise issues.

4.2 Training Graphs

Training and validation curves were analysed to study the model's learning pattern. The loss curves show smooth convergence with minimal divergence, indicating that the model is unable to suffer from overfitting. Similarly, we see a consistent improvement in performance over the training epochs and that the validation accuracy follows training very closely. Which we present as proof that our CNN architecture and preprocessing pipeline is highly optimized for this classification task.

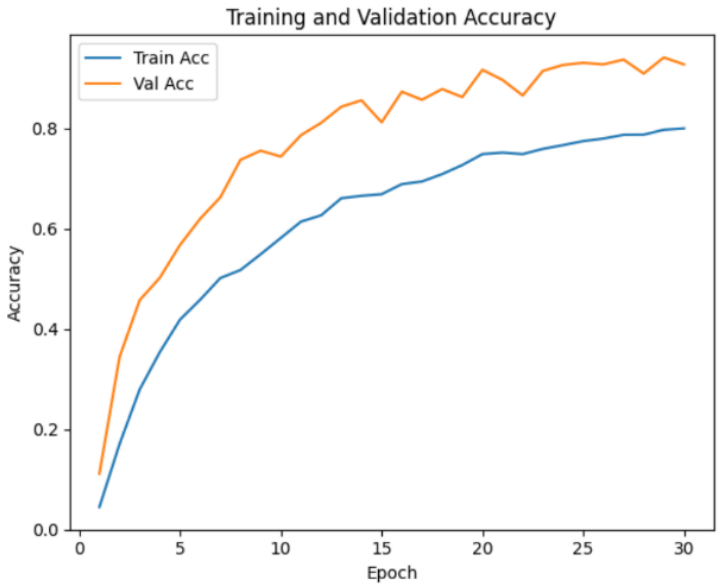


Fig.7.Accuracy Graph

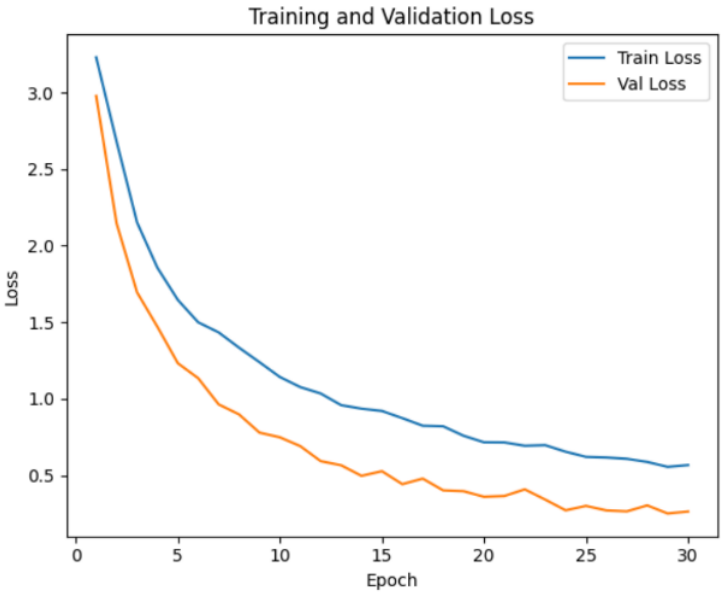


Fig.8.Loss Graph

5. Summary

This research presents a practical, low-cost system for recognizing ASL fingerspelling gestures and translate them into text and voice output. We integrated hand landmark extraction with CNN classification and also a GUI based user interaction which improves the system's communication access for Deaf and blind users. Also, we designed it for real time performance and high accuracy which makes it very much suited for real world applications.

6. Conclusion

Combining the efficiencies of real time ASL alphabet recognition and communication bridged all

gaps left by both deaf and hearing population. Through the application of skeleton-based landmark extractions along with a Kronecker neural network and a rule-based refinement, accuracy was maintained through real time processes. The user's accessibility was enhanced through converting the articulated speech and the text into speech with recognition. The system proved to accomplish all the challenges set by users with communication disabilities. Additionally, robust systems that utilize vision-based gesture recognition can be effective and seamless. The project is effective, lightweight, and is built on deep learning, preprocessing methodologies, and is designed to be user-oriented with the systems preferences.

7. Limitations

Although the system performs well, several limitations remain. First, it supports only static ASL alphabet gestures, while dynamic gestures such as “J” and “Z” are handled in a simplified manner. Second, the framework currently only accepts input as single-hand gestures which hinders the framework's effectiveness in the execution of more sophisticated sign languages consisting of two-hand interactions. Additionally, the training dataset utilized a rather limited group of signers, which may hinder the system's capability to generalize to various hand configurations and differing styles of signing. These deficiencies Establish the necessity for the collection of more diverse datasets and improvement of temporal modelling.

8. Future Scope

The employment of temporal neural architectures such as LSTMs, GRUs, and Transformers can facilitate the incorporation of dynamic gesture recognition. The incorporation of the ability to translate full-length sentences into the framework can be accomplished through the integration of NLP systems that can recognize and rectify grammar as well as comprehend meaning. System can also be used to support multi-hand gestures, enabling recognition of more complex signs. In future we may explore deploying on mobile devices for more accessibility for ASL users.

9. References

1. Pigou, L., Dieleman, S., Kindermans, P. J., & Schrauwen, B. (2014, September). Sign language recognition using convolutional neural networks. In *European Conference on Computer Vision* (pp. 572–578). Springer International Publishing.
2. Huang, J., Zhou, W., Li, H., & Li, W. (2015, June). Sign language recognition using 3D convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). IEEE.
3. Cui, R., Liu, H., & Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7361–7369).
4. Shah, J. A. (2018). Deepsign: A deep-learning architecture for sign language.
5. Taskiran, M., Killioglu, M., & Kahraman, N. (2018, July). A real-time system for recognition of American sign language by using deep learning. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)* (pp. 1–5). IEEE.
6. Abiyev, R. H., Arslan, M., & Idoko, J. B. (2020). Sign language translation using deep convolutional neural networks. *KSII Transactions on Internet & Information Systems, 14*(2).
7. Kadhim, R. A., & Khamees, M. (2020). A real-time American sign language recognition system using convolutional neural network for real datasets. *Tem Journal*, *9*(3), 937.

8. Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering, 28*(3).
9. Sundar, B., & Bagyammal, T. (2022). American sign language recognition for alphabets using MediaPipe and LSTM. *Procedia Computer Science, 215,* 642–651.
10. Jayanthi, P., Bhama, P. R. S., & Madhubalasri, B. (2023). Sign language recognition using deep CNN with normalised keyframe extraction and prediction using LSTM: Continuous sign language gesture recognition and prediction. *Journal of Scientific & Industrial Research (JSIR), 82*(07), 745–755.
11. Vyavahare, P., Dhawale, S., Takale, P., Koli, V., Kanawade, B., & Khonde, S. (2023). Detection and interpretation of Indian sign language using LSTM networks. *Journal of Intelligent Systems and Control, 2*(3), 132–142.
12. Alsharif, B., Altaher, A. S., Altaher, A., Ilyas, M., & Alalwany, E. (2023). Deep learning technology to recognize American sign language alphabet. *Sensors, 23*(18), 7970.
13. Triwijoyo, B. K., Karnaen, L. Y. R., & Adil, A. (2023). Deep learning approach for sign language recognition. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), 9*(1), 12–21.
14. Paul, S. K., Walid, M. A. A., Paul, R. R., Uddin, M. J., Rana, M. S., Devnath, M. K., … & Haque, M. M. (2024). An Adam based CNN and LSTM approach for sign language recognition in real time for deaf people. *Bulletin of Electrical Engineering and Informatics, 13*(1), 499–509.
15. Tao, T., Zhao, Y., Liu, T., & Zhu, J. (2024). Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges. *IEEE Access, 12,* 75034–75060.