International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(1) (April),2026
Received: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

SPEAKSIGNS: REAL-TIME GESTURE RECOGNITION AND SPEECH SYNTHESIS FOR SIGN LANGUAGE USERS

1. Dr. T. SRINIVASA RAO
Professor
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Andhra Pradesh, India
Email: srinivas123fast@gmail.com
2. SAYANI HARIKA
Student
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Andhra Pradesh, India
Email: harikasayani3@gmail.com
3. YERNINTI LOWSHIK TEJA
Student
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Andhra Pradesh, India
Email: lowshikteja@gmail.com
4. VANABATHINA SARITHA
Student
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Andhra Pradesh, India
Email: vsaritha810@gmail.com
5. USTELA HEMANTH NAGA KUMAR
Student
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Andhra Pradesh, India
Email: hemanthnagakumar21@gmail.com

ABSTRACT This paper presents a real-time Sign Language Recognition (SLR) framework capable of processing both static images and live video streams to generate corresponding textual representations and synthesized speech. The proposed system integrates Media Pipe based hand and poses tracking to achieve robust landmark extraction, followed by a hybrid deep learning architecture that combines Convolutional Neural Networks (CNNs) for spatial feature learning with Bidirectional Long Short-Term Memory (Bi- LSTM) networks for temporal sequence modeling. The framework supports both isolated sign recognition and continuous sign sentence translation, enabling efficient handling of dynamic gesture sequences. Additionally, a text-to-speech (TTS) module is incorporated to convert recognized text into naturalistic speech output, thereby facilitating real-time, bidirectional communication between deaf and hearing individuals. Extensive experimental evaluations conducted on multiple benchmark datasets demonstrate the effectiveness of the proposed approach, achieving up to 99% recognition accuracy along with high responsiveness and fluency. The results indicate the strong potential of the proposed framework for deployment in real-world assistive communication and human–computer interaction applications.

INDEX TERMS Sign Language Recognition (SLR), Convolutional Neural Networks (CNN), Bidirec- tional Long Short-Term Memory (Bi-LSTM), Media Pipe, Text-to-Speech (TTS), Assistive Technology, Human–Computer Interaction.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(1) (April),2026
Received: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

I. INTRODUCTION

Communication plays a fundamental role in human interaction, enabling individuals to express thoughts, emotions, and intentions. For individuals with partial or complete hearing loss, sign language serves as the primary medium of communication, relying on structured combinations of hand gestures, body movements, and facial expressions. While sign languages are linguistically rich and region-specific, effective communication between the signing community and hearing individuals remains a major challenge due to the lack of universal understanding of sign language [1], [2]. This communication gap significantly affects access to education, employment, healthcare, and social inclusion. Recent advances in computer vision and deep learning have led to increased research interest in Sign Language Recognition (SLR) systems aimed at bridging this gap [3], [4]. Traditional SLR approaches often relied on sensor-based devices such as data gloves, motion trackers, or depth sensors, which are expensive, intrusive, and unsuitable for real-world deployment. Vision-based methods using standard RGB cameras offer a cost-effective alternative; however, they introduce challenges such as background noise, hand occlusion, illumination vari- ations, signer dependency, and complex temporal dynamics in continuous gestures. The rapid evolution of deep neural networks particularly CNNs and RNNs has significantly im- proved the performance of vision-based SLR systems. CNNs have demonstrated strong capability in extracting spatial features from hand and pose images, while sequence models such as LSTM networks are effective in modeling temporal dependencies across gesture sequences [5]. However, many existing approaches are limited to isolated sign recognition, struggle with continuous sentence-level gestures, or lack real- time performance suitable for assistive applications. More- over, most current SLR systems focus solely on converting sign gestures into text, assuming that text output alone is sufficient for communication. In practical scenarios, speech output is essential for seamless interaction between deaf and hearing individuals. The integration of TTS systems with SLR frameworks remains underexplored, particularly in real-time environments that require low latency, high ac- curacy, and naturalistic output. To address these limitations, this paper proposes a real-time Sign Language Recognition framework capable of processing both static images and live video streams. The proposed system leverages MediaPipe-based hand and pose tracking to robustly extract spatial landmarks, reducing sensitivity to background variations and lighting conditions. Hybrid deep learning architecture is introduced, combining CNNs for spatial feature representation with Bi-LSTM networks for modeling temporal dependencies in continuous sign sequences. This design enables effective recognition of both isolated signs and continuous sign sentences. Furthermore, the framework incorporates a TTS module that converts recognized sign language text into synthesized speech, facilitating bidirectional and inclusive communication between deaf and hearing users. The entire system is designed for real-time operation, making it suitable for deployment in assistive technologies, human–computer interaction systems, and inclusive communication platforms. Extensive experiments conducted on multiple benchmark datasets demonstrate the effectiveness of the proposed approach, achieving recognition accuracies of up to 99%, along with high responsiveness and fluency in real-time scenarios. The results highlight the robustness, scalability, and practical applicability of the proposed framework for real-world assistive communication systems.

II. RELATED WORK

SLR aims to automatically interpret human hand gestures, body movements, and facial expressions to enable communication between deaf and hearing individuals. Although hand gestures play a dominant role, effective recognition requires the combined analysis of manual components (hand shape, orientation, motion) and non-manual components (facial expressions, head pose, and body posture). The diversity of sign languages across different regions, along with varia- tions in signing speed, signer appearance, and environmental conditions, makes real-time recognition a challenging task. Early SLR systems predominantly relied on sensor-based approaches, including data gloves, depth sensors, and motion capture devices. While these methods provided accu- rate gesture tracking, they suffered from high cost, limited portability, and user discomfort, restricting their applicability in real-world environments. To overcome these limitations, researchers increasingly adopted vision-based approaches using RGB cameras, which are non-intrusive and cost-effective but introduce challenges such as background clutter, illumination changes, occlusions, and viewpoint variations. With the advancement of deep learning, CNNs have become the foundation of modern SLR systems due to their strong ability to extract spatial features from images and video frames. Several studies have employed CNN-based architectures for isolated sign recognition, particularly for alphabet- and word-level gestures. However, such approaches often fail to generalize to continuous sign sequences, where temporal dependencies between gestures play a crucial role. To model temporal dynamics, Recurrent Neural Networks (RNNs), especially LSTM and Bi-LSTM networks, have been widely explored [3], [6], [7]. Hybrid architectures combining CNNs for spatial feature extraction and LSTM-based models for temporal modeling have demonstrated improved performance in continuous sign recognition tasks. These models capture both short-term and long-term dependencies across gesture sequences, enabling more accurate sentence-level recognition. Despite these improvements, many ex-

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(1) (April),2026
Received: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

isting frameworks struggle with real-time performance and robustness across different signers and environments. Recent research has shown that landmark-based representations can significantly enhance SLR performance. Frameworks utilizing pose and hand keypoints reduce sensitivity to background noise and lighting variations while preserving essential motion information. In particular, MediaPipe-based hand and pose tracking has gained attention due to its real-time capability, high accuracy, and robustness in unconstrained environments. Several studies have incorporated skeletal landmarks with deep learning models, demonstrating improved generalization and computational efficiency compared to raw image-based methods [8], [9]. Another limitation in existing SLR systems is their focus on text-based output. While converting sign language into text is useful, it does not fully support natural interaction between deaf and hearing individuals. TTS integration remains relatively unexplored in SLR literature, especially in real-time systems. The absence of speech output limits the usability of SLR systems in assistive communication scenarios, where auditory feedback is essential for hearing users. Although significant progress has been made in recognizing American Sign Language (ASL), Chinese Sign Language (CSL), and other widely studied sign languages, many systems lack scalability and real-time responsiveness, particularly for continuous sign sentence recognition. Furthermore, performance degradation is commonly observed under real-world conditions involving varying signer styles, occlusions, and dynamic backgrounds. To address these gaps, the proposed work introduces a real-time SLR framework that integrates MediaPipe-based landmark extraction with a hybrid CNN–Bi-LSTM architecture for robust spatial–temporal modeling. Additionally, the inclusion of a Text-to-Speech module enables natural spoken output, facilitating seamless bidirectional communication. The proposed system is designed to operate efficiently in real-world settings while supporting both isolated and continuous sign recognition.

III. PROPOSED METHODOLOGY

A. SYSTEM OVERVIEW

The proposed Sign Language Recognition (SLR) framework is designed to perform real-time recognition and translation

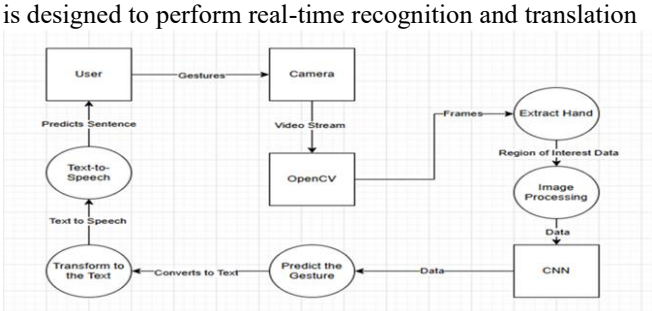

FIGURE 1. Architecture of Sign Language Recognition System

of sign language using both static images and live video streams. The system follows an end-to-end pipeline that captures visual sign inputs, extracts discriminative spatial and temporal features, and converts the recognized gestures into textual and speech outputs. MediaPipe is employed for accurate hand and pose landmark detection, which enables reliable representation of sign gestures while minimizing the impact of background noise, lighting variations, and signer appearance. CNN and Bi-LSTM is used to model spatial and temporal characteristics of sign language gestures. The final recognized text is converted into synthesized speech using a TTS module, facilitating seamless communication between deaf and hearing individuals. The framework operates using standard RGB cameras and is optimized for low latency and real-time performance, making it suitable for assistive communication and human–computer interaction applications.

B. INPUT ACQUISITION

The input to the proposed framework consists of either static sign images or continuous live video streams. Static images are processed independently and are primarily used for isolated sign recognition. Live video input is captured using a standard webcam and segmented into individual frames at fixed time intervals to preserve temporal continuity. Each frame undergoes preprocessing operations including resizing, normalization, and noise reduction to ensure consistency across inputs. These preprocessing steps help mitigate variations in resolution, illumination, and camera positioning. By supporting both static and dynamic inputs, the proposed system provides flexibility for offline sign analysis as well as real-time sign language interpretation. [10], [11]

C. HAND AND POSE LANDMARK EXTRACTION USING MEDIAPIPE

Accurate extraction of hand and pose landmarks is essential for recognizing the fine-grained articulations involved in sign language. In the proposed framework, MediaPipe is used to detect and track key landmarks corresponding to hand joints, finger positions, wrist orientation, and upper-body pose [12]. Landmark extraction is performed on each image or video frame independently, ensuring consistent detection across static and dynamic inputs. The extracted landmark coordinates are normalized with respect to frame dimensions to achieve scale invariance and robustness to camera distance. This landmark-based representation significantly reduces dependency on background information and lighting conditions while preserving essential structural and motion-related features. The resulting landmark vectors provide a compact and informative representation of sign gestures, which is well suited for deep learning-based feature extraction.

D. SPATIAL FEATURE EXTRACTION USING CONVOLUTIONAL NEURAL NETWORKS

Spatial features of sign gestures are learned using a CNN. It processes the preprocessed frames along with the landmark-enhanced representations to capture discriminative spatial patterns such as hand shape, finger configuration, and rel-

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(1) (April),2026
Received: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

ative joint positioning. Convolutional layers apply multiple filters to extract local spatial features, while pooling lay- ers reduce dimensionality and enhance robustness to minor spatial variations. Through hierarchical feature learning, the CNN captures both low-level features such as edges and contours and high-level semantic representations of gestures. Extracted feature maps are reshaped into a single vector and then processed by fully connected layers to produce compact feature representations, which serve as effective spatial representations for subsequent temporal modeling.

E. TEMPORAL MODELING USING BIDIRECTIONAL LSTM

To effectively recognize continuous sign language sequences, temporal dependencies across consecutive frames must be modeled [13], [14]. The proposed framework employs a Bi-LSTM network to analyze temporal relationships in both forward and backward directions. The CNN-extracted feature vectors are sequentially fed into the Bi-LSTM net- work, allowing the model to capture contextual information from preceding and succeeding frames simultaneously. This bidirectional modeling improves recognition accuracy by effectively handling gesture transitions, motion continuity, and variations in signing speed. The memory mechanism of LSTM cells enables learning of long-term dependencies and mitigates the vanishing gradient problem. The output of the Bi-LSTM network is passed through dense layers with a softmax activation function to generate the final textual prediction corresponding to the input sign sequence.

F. TEXT-TO-SPEECH CONVERSION

To enhance accessibility and enable effective communication with hearing individuals, the recognized text output is converted into speech using a Text-to-Speech module [15], [16]. The TTS component synthesizes natural-sounding speech in real time based on the generated textual representation. This auditory output allows immediate feedback and improves usability in real-world assistive communication scenarios. The integration of text and speech output makes the system suitable for environments where visual interpretation alone may not be sufficient.

G. DATASET DESCRIPTION

The proposed system is trained and evaluated using pub- licly available benchmark datasets, including the Word-Level American Sign Language (WLASL) dataset and standard ASL datasets [17]. WLASL dataset consists of a large collection of word-level sign videos recorded from multi- ple signers under diverse recording conditions, providing significant variability in gesture execution, viewpoints, and motion dynamics. This dataset is particularly suitable for evaluating both isolated sign recognition and continuous sign sentence translation. In addition, ASL datasets containing static images and video sequences are used to assess the generalization capability of the proposed framework. These datasets include variations in signer appearance, background settings, and signing speed, making them representative of real-world scenarios. Rotation, scaling, and temporal sampling are applied during training to improve robustness and reduce overfitting. The use of WLASL and ASL datasets ensures comprehensive evaluation of the system across both static and dynamic sign language recognition tasks.

FIGURE 2. Sample dataset images

Sample images from the dataset used in this study are shown in Figure 2.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental evaluation of the proposed real-time SLR framework developed using MediaPipe-based landmark extraction and a hybrid CNN–BiLSTM architecture. The experiments are conducted to validate the effectiveness of the proposed approach for both isolated sign recognition and continuous sign sentence recognition using benchmark datasets. Performance is analyzed using standard evaluation metrics, and results demonstrate the robustness and accuracy of the proposed framework in real-world scenarios.

A. EXPERIMENTAL SETUP

All experiments are conducted on a GPU-enabled workstation to ensure efficient training and real-time inference. The proposed model is implemented using Python with deep learning frameworks, and MediaPipe is used for real-time hand and pose landmark extraction. The system is evaluated on the WLASL dataset and ASL datasets, which include both static images and continuous video sequences. The datasets are divided into training, validation, and testing sets following standard protocols to ensure unbiased evaluation. For isolated sign recognition, approximately 25,000 images are used for training and 5,000 images are used for validation and testing. For continuous sign recognition, video sequences are segmented into frame-level inputs, and temporal features are learned using the Bi-LSTM network. Data augmentation techniques such as rotation, scaling, and temporal sampling are applied to improve model generalization and reduce overfitting.

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(1) (April),2026
Received: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

B. PERFORMANCE EVALUATION OF CNNBILSTM MODEL

The recognition performance of the proposed hybrid CNN–BiLSTM model is evaluated using accuracy and loss curves during training and validation phases. Convergence behavior of the model indicates stable learning and effective feature representation. The proposed model achieves con- sistent improvement in recognition accuracy with reduced validation loss, demonstrating its capability to learn discrimi- native spatial and temporal features from sign language data. To further analyze classification performance, a confusion matrix is generated for the test dataset. The confusion matrix illustrates strong diagonal dominance, indicating high true positive rates across most sign classes and minimal misclas- sification between visually similar gestures. These results confirm the effectiveness of the CNN for spatial feature ex- traction and the Bi-LSTM for temporal modeling of gesture sequences.

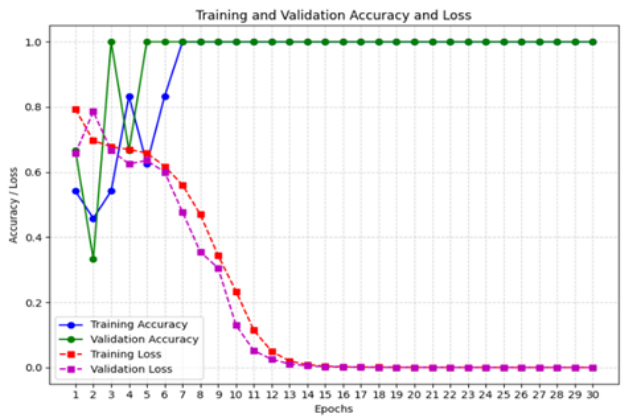

FIGURE 3. Training and validation accuracy/loss over 5 epochs.

Above figure shows rapid improvement in training ac- curacy, reaching near-perfect performance, while validation accuracy peaks early and then declines, indicating overfitting and limited generalization beyond initial epochs.

C. EVALUATION METRICS

The performance of the proposed model is quantitatively evaluated using standard classification metrics including Pre- cision, Recall, F1-score, and Accuracy. These metrics are computed based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

Recall is computed as

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

The F1-score, which represents the harmonic mean of Preci- sion and Recall, is calculated as

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Accuracy is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

The computed results show that the proposed CNN–BiLSTM framework achieves recognition accuracy of up to 99% on the WLASL dataset and demonstrates comparable performance on ASL datasets. These results significantly outperform sev- eral existing sign language recognition approaches, particu- larly in continuous sign recognition scenarios.

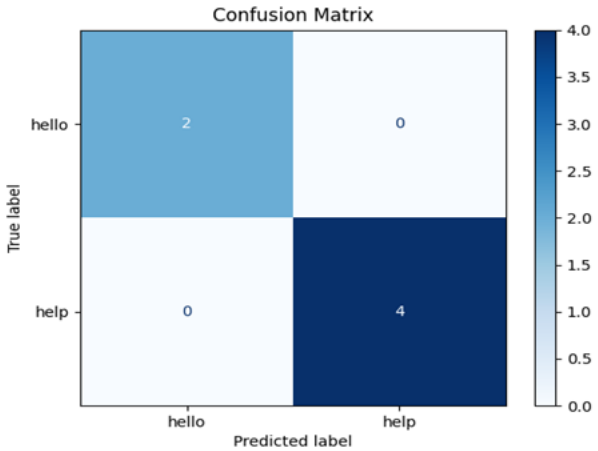

FIGURE 4. Sample Confusion Matrix

The figure illustrates the classification performance of the hybrid CNN-BiLSTM model on a small subset of words, showing perfect recognition for the word “help” and accurate recognition for “hello.”

D. COMPARATIVE ANALYSIS

A comparative analysis is conducted between the proposed framework and existing sign language recognition models reported in the literature. The comparison highlights the advantages of integrating MediaPipe-based landmark extrac- tion with deep spatial–temporal modeling. Unlike traditional image-based approaches, the proposed method effectively handles background variations, signer diversity, and tempo- ral dependencies. The improved recognition accuracy and reduced misclassification rates validate the superiority of the proposed approach over conventional CNN-only or LSTM- only models.

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(1) (April),2026
Received: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

E. REAL-TIME PERFORMANCE ANALYSIS

In addition to recognition accuracy, real-time performance is evaluated to assess the suitability of the proposed system for practical deployment. The framework demonstrates low inference latency and high frame processing rates, enabling real-time recognition of live sign language input using a standard webcam. The integration of MediaPipe significantly reduces computational overhead by focusing on landmark-based representations rather than raw pixel-level processing. This ensures efficient execution without requiring specialized sensors or hardware.

F. TEXT-TO-SPEECH OUTPUT EVALUATION

The recognized text output is converted into synthesized speech using a Text-to-Speech module to facilitate communication with hearing individuals. The TTS component operates in real time and produces intelligible and natural-sounding speech. The integration of speech output enhances the usability of the system in assistive communication environments such as classrooms, public service centers, and human–computer interaction systems.

FIGURE 5. Sample UI application results for sign language recognition using live webcam

The interface captures hand gestures via webcam and displays the predicted word, last accepted word, and constructed sentence in real time

G. DISCUSSION

The proposed real-time SLR framework effectively addresses key challenges in sign language interpretation, including temporal variation, signer independence, and environmen- tal complexity. The hybrid CNN–BiLSTM architecture successfully captures both spatial and temporal features, while MediaPipe-based landmark extraction improves robustness and efficiency. The high recognition accuracy achieved on WLASL and ASL datasets confirms the reliability and scalability of the proposed system for real-world applications.

V. CONCLUSION

This paper presented a real-time SLR framework capable of processing both static images and live video streams to generate corresponding textual output and synthesized speech. The proposed system integrates MediaPipe-based hand and pose landmark extraction with CNN + Bi-LSTM for spatial feature learning and for temporal sequence modeling. This combination enables effective recognition of isolated signs as well as continuous sign sequences, addressing key challenges such as temporal dependency, signer variation, and real-time performance. The framework was experimentally evaluated using benchmark datasets, including WLASL and standard ASL datasets. Extensive experimental results demonstrate that the proposed CNN–BiLSTM model achieves high recognition accuracy, reaching up to 99%, while maintaining low inference latency suitable for real-time applications. The use

5

of MediaPipe significantly enhances robustness by enabling accurate landmark-based feature extraction, thereby reducing

6

sensitivity to background noise and illumination variations. Furthermore, the incorporation of a TTS module allows the recognized text to be converted into natural and intelligible speech, facilitating effective communication between deaf and hearing individuals. Overall, the results confirm that the proposed framework provides an efficient, scalable, and practical solution for real-world sign language interpretation and assistive communication systems. Future work will focus on extending the framework to support larger vocabularies, multilingual sign languages, and sentence-level grammatical restructuring. Additionally, integrating transformer-based temporal models and improving signer-independent recognition will further enhance the system's performance and applicability in diverse real-world environments.

REFERENCES

[1] A. S. Cheok, R. S. S. Teh, and M. H. Tan, "Vision-based hand gesture recognition: A review," International Journal of Human-Computer Studies, vol. 79, pp. 65–79, 2015.
[2] A. Karpagavalli and A. Chandra, "A review on automatic sign language recognition," International Journal of Engineering and Technology, vol. 5, no. 2, pp. 196–201, 2013.
[3] C. K. Lee, K. K. Ng, C. H. Chen, H. C. Lau, and T. Tsoi, "American sign language recognition and training method with recurrent neural networks," Expert Systems with Applications, vol. 167, April 2021.
[4] S. Aly and W. Aly, "Deeparslr: A signer-independent deep learning framework for isolated arabic sign language recognition," IEEE Access, vol. 8, pp. 83 199–83 212, 2020.
[5] Keerthi Guttikonda, Yerninti Ashvitha, Velagala Sai Ranga Reddy, Ramba Murali Krishna, and Penumala Sandeep, "Integrating convolutional neural networks and machine learning for ac- curate identification of autism spectrum disorder using facial biomarkers," in Proceedings of the 2024 IEEE International Conference on Emerging Systems and Intelligent Computing (ESIC). IEEE, Feb. 2024.
[6] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2015, pp. 2625–2634.
[7] J. Zhao, W. Shi, and Y. Guo, "Sign language recognition using deep cnn and bi-lstm," Multimedia Tools and Applications, vol. 80, pp. 1–19, 2021.

[8] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," Expert Systems with Applications, vol. 150, July 2020.
[9] P. Molchanov et al., "Hand gesture recognition with 3d convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 2015, pp. 1–7.
[10] Keerthi Guttikonda, G Ramachandran, and GVSNRV Prasad, "Autism spectrum disorder prediction using lasso regularised bat search optimisation," In- ternational Journal of Services Operations and Informatics, 2024.
[11] Dr Guttikonda Keerthi, Dr G Ramachandran, Dr GVSNRV Prasad, "Cuckoo search optimization-based feature selection for predicting autism spectrum disorder using artificial immune algorithms," Journal of Theoretical and Applied Information Technology, 2025.
[12] Allada Koteswaramma, M. Babu Rao, and G. Jaya Suma, "An intelligent adaptive learning framework for fake video detection using spatiotemporal features," Signal, Image and Video Processing, 2024, cited by 4.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[14] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm networks," in Proceedings of the IEEE International Joint Conference on Neural Networks, 2005, pp. 2047–2052.
[15] P. Taylor, Text-to-Speech Synthesis. Cambridge University Press, 2009.
[16] A. Black and K. Tokuda, "Statistical parametric speech synthesis," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2005, pp. 1229–1232.
[17] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods compari- son," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, March 2020, pp. 1459–1469.