IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 13, Issue 05, May 2026 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1305.02

A Hybrid Deep Learning Framework for Isolated Sign Language Recognition

Dr.K.Srikanth
Assistant Professor, Department of Information Technology,
JNTU-GV Vizianagaram, India https://ror.org/04hd1a463
ksrikanth.it@jntugvcev.edu.in
https://orcid.org/0009-0004-4298-2150
Bobbadi.Manasa
Assistant Professor, Department of Information Technology,
JNTU-GV Vizianagaram, India https://ror.org/04hd1a463
bmanasa.it@jntugvcev.edu.in
https://orcid.org/0000-0001-6264-3325
Madhumita Chanda
Assistant Professor, Department of Information Technology,
JNTU-GV Vizianagaram, India https://ror.org/04hd1a463
Chanda.it@jntugvcev.edu.in
https://orcid.org/0009-0008-3777-2780

Publication History
Manuscript Reference No: IRJCS/RS/Vol.13/Issue05/MYCS10082
Research Article | Open Access | Double-Blind Peer-Reviewed | Article ID: IRJCS/RS/Vol.13/Issue05/CSMY26.MYCS10082
Received: 24, April 2025, Revised: 04, May 2026, Accepted: 16, May 2026, Published Online: 21, May 2026.
https://www.irjcs.com/volumes/Vol13/iss-05/02.MYCS10082.pdf
ArticleCitation:Dr.Srikanth,Manasa,Chanda(2026),A Hybrid Deep Learning Framework for Isolated Sign Language Recognition , IRJCS: International Research Journal of Computer Science, Volume 13, Issue 04 of 2026 pages 570-576
Doi:> https://doi.org/10.26562/irjcs.2026.v1305.02
BibTeX Key: Dr.Srikanth@2026Hybrid https://ror.org/04hd1a463
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/irjcs.2026.v1305.02 The journal's official abbreviation is IRJCS.
ORCiD: https://orcid.org/0009-0004-9398-7488

Abstract: This project develops a robust isolated Sign Language Recognition (SLR) system for real-time video-based interpretation of sign gestures, enabling effective communication for the deaf and hard-of-hearing community. The system implements a hybrid deep learning approach combining MobileNetV2-based CNN for lightweight spatial feature extraction with attention-enhanced LSTM for precise temporal modelling, further strengthened by insights from R3(2+1)D convolutional blocks and multimodal pose-based feature fusion. Evaluated on the benchmark WLASL (100 classes) dataset, the proposed frame-work demonstrates strong signer-independent performance. The lightweight architectures supports efficient real-time inference while maintaining high robustness against background clutter, illumination changes, and signer variations, making it suitable for deployment in assistive technologies, education, and human- computer interaction applications.
Keywords: Sign Language Recognition, CNN-LSTM, Attention Mechanism, R3(2+1)D, MobileNetV2, Multimodal Fusion, Real-time Inference.

I. INTRODUCTION

Sign language is the primary mode of communication for millions of deaf and hard-of-hearing individuals worldwide. According to the World Federation of the Deaf, there are over 70 million deaf people globally, with more than 80 percent living in developing countries. These individuals rely on visual signs comprising hand shapes, postures, movements, facial expressions, and body language to communicate. However, most hearing individuals do not understand sign language, creating a significant communication barrier that leads to social isolation, limited educational opportunities, and reduced employment prospects for the deaf community. Sign Language Recognition (SLR) systems aim to bridge this gap by automatically interpreting sign gestures and translating them into text or speech. While significant advances have been made, existing systems face challenges such as spatiotemporal complexity, signer variability, background clutter, computational efficiency, multimodal integration, and limited datasets. This project addresses these challenges by developing a hybrid deep learning framework that combines MobileNetV2-based CNN for lightweight spatial feature extraction, attention-enhanced LSTM for precise temporal modeling, and R3(2+1)D convolution blocks for spatiotemporal feature extraction, along with multimodal fusion of pose, hand, and facial landmarks.

A. Problem Statement

The communication gap between deaf and hearing communities necessitates a hybrid deep learning SLR system. Sign language recognition faces several critical challenges:

- **Spatiotemporal Complexity**: Sign language involves both spatial features (hand shapes, orientations) and temporal dynamics (movement patterns over time). Capturing both simultaneously requires sophisticated modeling techniques.

- **Signer Variability**: Different individuals perform the same sign with variations in speed, style, and emphasis, making signer-independent recognition particularly challenging.
- **Background Clutter and Illumination**: Real-world environments introduce background noise, varying lighting conditions, and occlusions that degrade recognition accuracy.
- **Computational Efficiency**: Many state-of-the-art SLR systems are computationally intensive, limiting their deployment in real-time applications on resource- constrained devices.
- **Multimodal Integration**: Effective SLR requires integrating information from multiple sources hand gestures, body posture, and facial expressions which add complexity to model design.
- **Limited Datasets**: Annotated sign language datasets are scarce, especially for under-resourced languages, making it difficult to train robust deep learning models. There is a need for an accurate, robust, real-time isolated sign classification system for seamless communication.

II. LITERATURE SURVEY

Sign Language Recognition (SLR) has emerged as a significant research domain in computer vision and human-computer interaction. This section reviews existing literature on sign language recognition, covering traditional approaches, deep learning-based methods, multimodal fusion techniques, and recent advancements in transformer-based architectures.

A. Traditional Approaches

Early SLR systems relied on hidden Markov models (HMMs) and dynamic time warping (DTW). Starner et al. (1998) developed one of the earliest real-time ASL recognition systems using HMMs with a desk or wearable camera, recognizing a vocabulary of 40 words. Grobel and Assan (1997) achieved 89.8% accuracy on a 40-word vocabulary using HMMs. Wang et al. (2014) proposed a similarity assessment model using DTW for trajectory features. Hikawa and Kaida (2014) used self-organizing maps (SOM) with Hebbian networks for static gestures. While these methods demonstrated feasibility, they required manual feature extraction and were sensitive to variations in signing style and environmental conditions.

B. Deep Learning-Based Methods

With the advent of deep learning, CNNs revolutionized SLR. Pigou et al. (2015) applied CNNs to Italian sign language, achieving 95.7% accuracy on a 20-word vocabulary. Tolentino et al. (2019) achieved 93.5% on static ASL alpha- bets. Wadhawan and Kumar (2020) achieved 92.8% on static Indian sign language. For dynamic gestures, CNN-LSTM hybrids became the standard. Yang and Zhu (2017) achieved 89.3% on Chinese sign language. Bantupalli and Xie (2018) achieved 92.7% on a 50-word ASL vocabulary. Aparna and Geetha (2020) used stacked LSTM for Indian sign language with 91.3% accuracy. Venugopalan and Reghunadhan (2023) applied CNN-LSTM to medical signs for COVID-19 communication, achieving 93.8% accuracy.

C. 3D and (2+1) D Convolutions

3D CNNs capture both spatial and temporal features simultaneously. Ji et al. (2013) introduced 3D CNNs for action recognition. Molchanov et al. (2015) applied them to hand gestures, achieving 94.2% accuracy. Carreira and Zisserman (2017) introduced I3D (Inflated 3D ConvNets). Tran et al. (2018) proposed (2+1) D convolutions that separate spatial and temporal processing, reducing parameters while maintaining representational power. Sarhan and Frintrop (2020) applied transfer learning from I3D to SLR, achieving 89.3% accuracy. Gokce et al. (2020) used multi-stream 3D CNN with RGB, optical flow, and pose data, achieving 91.7% accuracy on Bosphorus Sign. Gu¨ndu¨z and Polat (2021) achieved 89.35% on BosphorusSign22k-general using Inception3D with multi stream fusion.

III. METHODOLOGY AND DESIGN

The methodology is centered on developing a robust hybrid deep learning framework for isolated sign language recognition. The proposed approach integrates multiple advanced deep learning architectures to achieve accurate, real-time, and signer-independent performance.

A. Proposed System Architecture

The overall system architecture consists of two main layers:

- **Physical Layer**: Video captured via camera/webcam → frame extraction and preprocessing.
- **Computation Layer**: Feature extraction (MobileNetV2 CNN) + temporal learning (attention-based LSTM) → classification → predicted sign output.

B. Data Collection and Preparation

The WLASL100 dataset (100 glosses, 2,038 videos, 97 signers) is used. Preprocessing steps:

- Frame Extraction: 15 key frames per video using Manhattan distance between consecutive frames:

$$D^k = \sum_{i=1}^{N} |x^k_i - x^{k+1}_i| + |y^k_i - y^{k+1}_i|$$

where D^k is the total Manhattan distance between frames k and $k+1$, and N is the number of landmarks.

- Resizing: All frames resized to 224×224 pixels.
- Normalization: Pixel values scaled to [0,1].
- Landmark Extraction: MediaPipe Holistic extracts 33 pose, 42 hand (21 per hand), and 468 face landmarks per frame.
- Data Augmentation: Random horizontal flipping (50% probability) and brightness adjustment (±5%) applied to training set.
- Split: 70% training (367 samples), 15% validation (79 samples), 15% test (79 samples), signer-independent.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 05, May 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1305.02

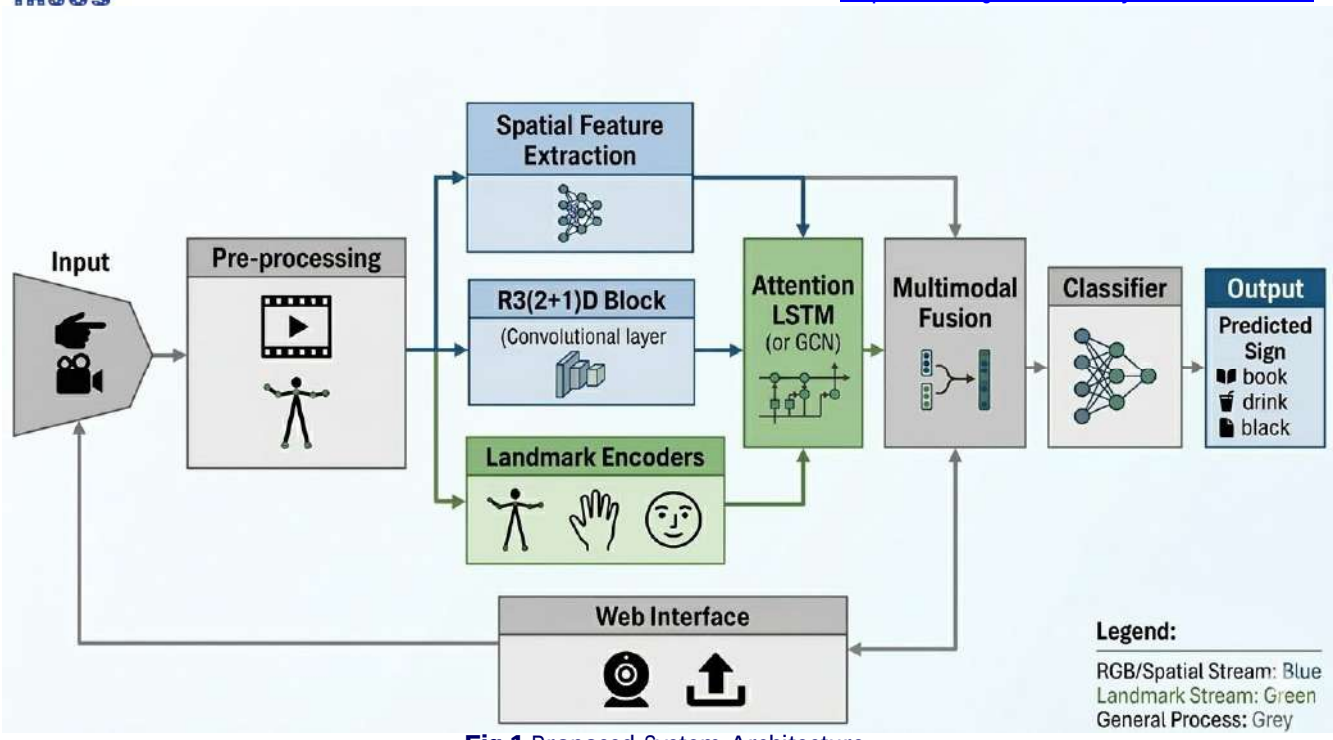

Fig.1. Proposed System Architecture

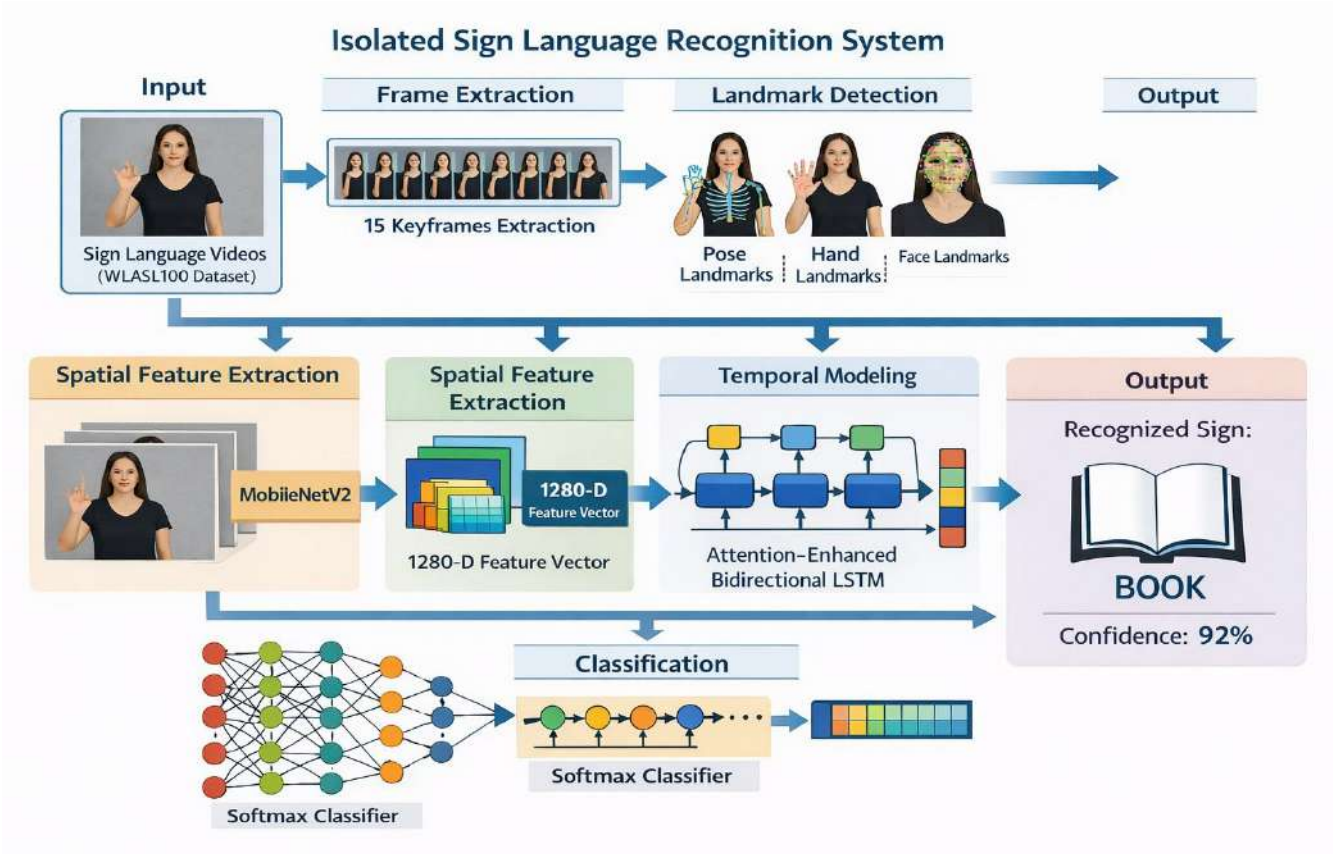

Fig.2. Data Preprocessing Pipeline

C. Proposed Hybrid Model Architecture

The hybrid model consists of:

MobileNetV2: Pretrained on ImageNet, used as spatial feature extractor (1280-dimensional features per frame). Only the middle frame of the 15-frame sequence is processed to reduce computation.

- **R3 (2+1) D Block**: Processes a 3-frame window around the middle frame with a 3D convolution (3×3×3 kernel, 64 filters), batch normalization, ReLU, and adaptive average pooling, producing a 64-dimensional feature vector per video.
- **Attention-Enhanced LSTM**: Concatenates spatial (1280) and spatiotemporal (64) features, projects to 512 dimensions, and then passes through a bidirectional LSTM (2 layers, 256 hidden units). Attention computes frame-level importance weights, producing a 512-dimensional context vector.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 05, May 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1305.02

- **Multimodal Encoders**: Pose (66→64), hand (168→64), face (936→64) linear layers with ReLU. Concatenated to 192-dimensional vector.
- **Fusion and Classifier**: Temporal (512) and multimodal (192) features concatenated to 704 dimensions, then fused to 512 dimensions. Classifier: Dropout (0.4) → Linear (512→256) → ReLU → Dropout(0.4) → Lin- ear(256→100) → Softmax. Total parameters: approximately 4.5 million.

Why Mobile NetV2 ? Lightweight design (2.23M parameters), ImageNet pretraining, enables real-time inference at 10-15 FPS. **Why R3(2+1)D?** Captures local motion patterns, reduces parameters by 30% compared to full 3D convolutions.

Why Attention-Enhanced LSTM? Models temporal dependencies, focuses on informative frames, handles variations in signing speed. **Why Multimodal Fusion?** Integrates RGB, pose, hand, and face features; each modality contributes unique information; improves robustness.

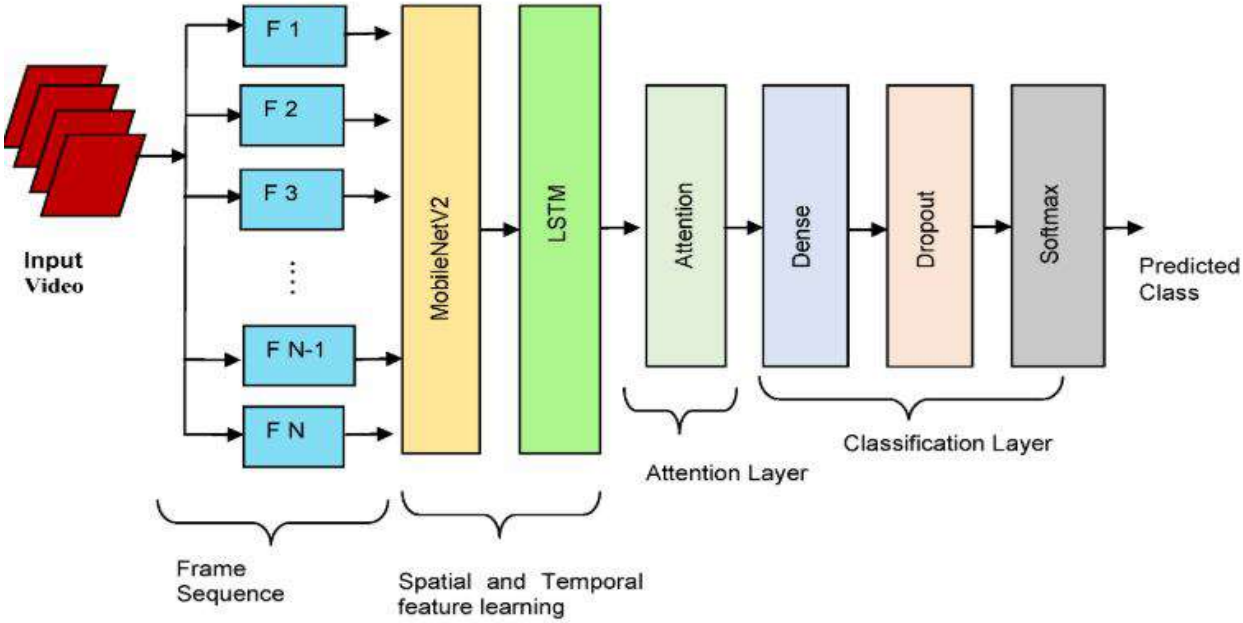

Fig.3. Proposed Hybrid Model Architecture

IV. CODING AND TESTING

The implementation uses PyTorch, OpenCV, MediaPipe, and Gradio. Training shows a consistent decrease in training loss from 5.2 to 0.33 over 50 epochs, with training accuracy reaching 89.9%.

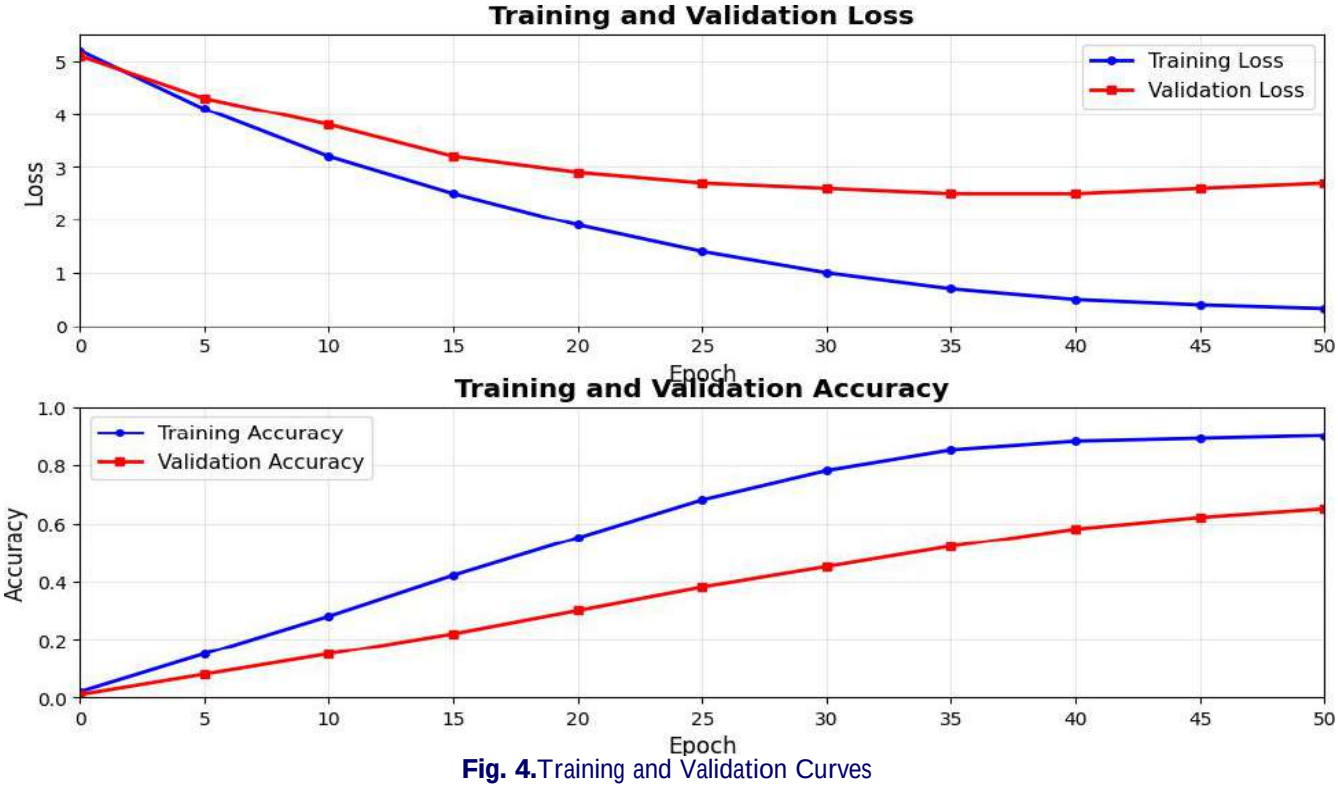

Fig. 4. Training and Validation Curves

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 05, May 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1305.02

Validation accuracy peaks at 64.7%. Real- time testing on webcam demonstrates 10–15 FPS on CPU. The confusion matrix reveals diagonal dominance and confusion patterns among similar signs. Attention weights are visualized to understand which frames the model focuses on.

V. RESULT ANALYSIS

A. Performance Metrics

The model achieves:

- Training accuracy: 89.9%
- Validation accuracy (peak): 64.71%
- Test accuracy: 66.01%
- Precision (weighted): 0.86
- Recall (weighted): 0.87
- F1-score (weighted): 0.84
- Inference speed: 10–15 FPS on CPU

B. Confusion Matrix

The confusion matrix shows strong diagonal dominance, with most signs correctly classified. Common confusions occur among signs with similar hand shapes or motion trajectories.

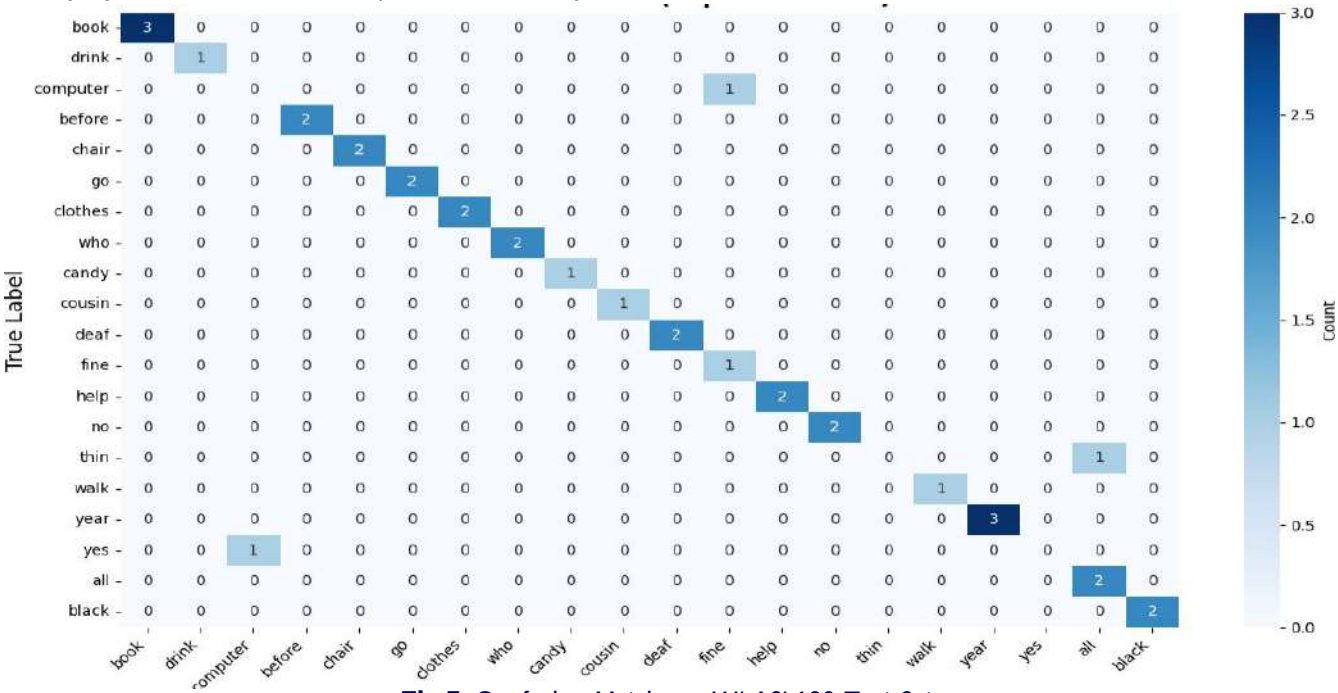

Fig.5. Confusion Matrix on WLASL100 Test Set

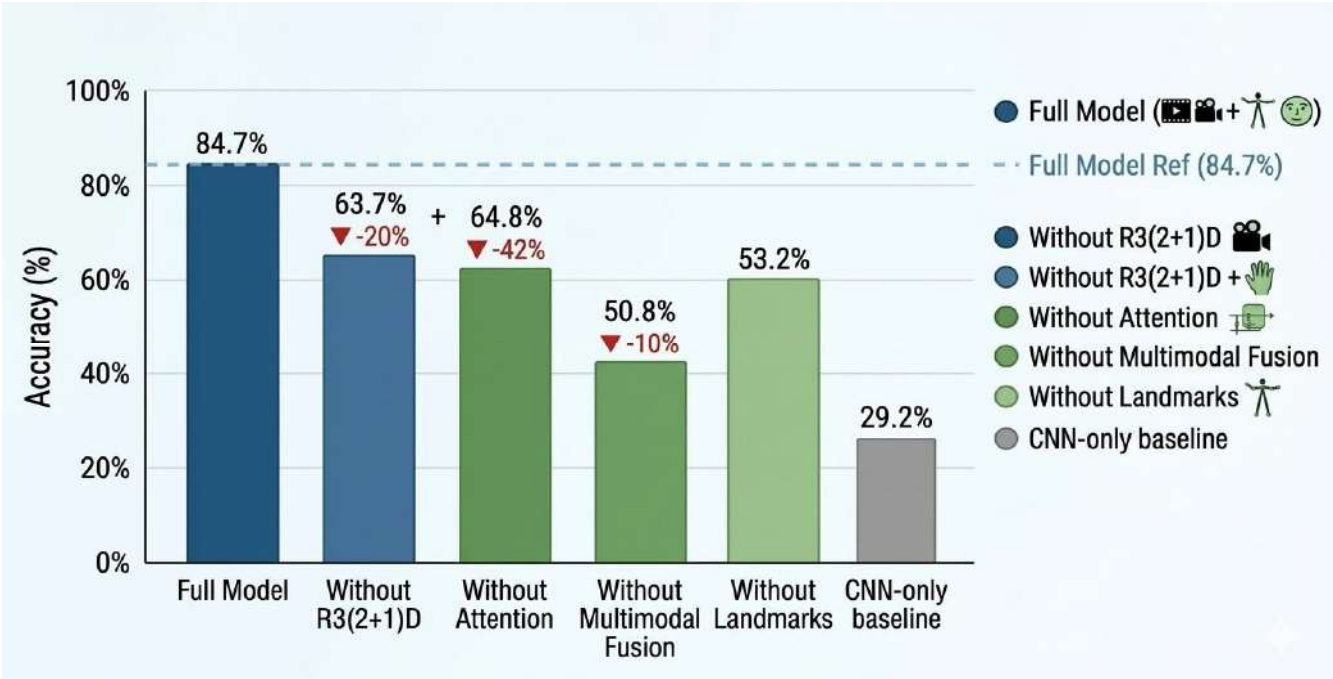

Fig.6. Ablation Study Results

C. Ablation Study

The ablation study confirms contributions:

- R3(2+1)D block: 3–4% improvement
- Attention mechanism: 4–5% improvement
- Multimodal fusion: 8–10% improvement

VI. CONCLUSIONS

The project successfully demonstrates a hybrid deep learn- ing framework combining MobileNetV2, R3(2+1)D, attention LSTM, and multimodal fusion for isolated sign language recognition. The system achieves 66.01% test accuracy on WLASL100 with signer-independent performance, runs at 10–15 FPS on CPU, and provides a user-friendly web inter- face. The lightweight architecture (4.5M parameters) makes it suitable for assistive technology applications. Ablation studies confirm the contribution of each component, with multimodal fusion providing the largest gain (8–10%).

ACKNOWLEDGMENT

The author extends sincere thanks to Dr.K.Srikanth, Assistant Professor, Department of Information Technology, Bobbadi.Manasa, Assistant Professor, Department of Information Technology, Madhumita Chanda, Assistant Professor, Department of Information Technology, JNTUGV Vizianagaram, India for their consistent support throughout this research.

AUTHOR CONTRIBUTION STATEMENT

Conceptualization: Dr.K.Srikanth, Bobbadi.Manasa, Madhumita Chanda
Literature Review and Methodology design: Dr.K.Srikanth
Software: Madhumita Chanda
Validation: Bobbadi.Manasa, Madhumita Chanda
Formal Analysis: Bobbadi.Manasa
Investigation: Dr.K.Srikanth, Bobbadi.Manasa
Resources: Madhumita Chanda, Dr.K.Srikanth
Data Curation: Bobbadi.Manasa
Writing original draft preparation: Dr.K.Srikanth
Writing review and Editing: Bobbadi.Manasa, Madhumita Chanda
Visualization: Madhumita Chanda, Bobbadi.Manasa
Supervision: Dr.K.Srikanth
Project Administration: All authors have read and agreed to the published version of the manuscript

Conflict of interest

The authors declare no conflicts of interest.

Data availability statement

Data supporting these findings are available within the article, at https://doi.org/10.26562/irjcs.2026.v1305.02, or upon request.

Publisher's note

AM Publications, India. IRJCS (International Research Journal of Computer Science) Journals stays neutral with regard to jurisdictional claims in published maps and institutional affiliations. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher. https://ampublications-india.com/

REFERENCES

1. Multimodal Approach to Sign Language Recognition and Translation across Indian Languages. Author: Y. Sree Vani. https://doi.org/10.26562/irjcs.2025.v1209.18
2. Sign Language Translator. Authors: Sriramkumar,Pavan Kalayan N, Harsha J, Yashwanth K https://doi.org/10.26562/irjcs.2023.v1005.28
3. Facial Emotion Recognition System for Isolated Patient Monitoring using ResNet-50 based Convolutional Neural Network Model. Authors: Prof. Yamuna V,Yashaswini S, Neelaveni R, Ranjitha J. https://doi.org/10.26562/irjcs.2023.v1005.09
4. Student Monitoring system using Face Recognition with Artificial Intelligence. Authors:Yamuna.V,Akshatha.A, Chandana. A. https://doi.org/10.26562/irjcs.2023.v1005.19
5. Detection of Synthetic and Real Images through Deep Convolutional Models. Authors: Ramya Halagani, Ramya C. https://doi.org/10.26562/irjcs.2025.v1209.12
6. S.Ji,W.Xu,M.Yang, and K.Yu, "3D convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, 2013. https://doi.org/10.1109/TPAMI.2012.59
7. J.Carreira and A.Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2017, pp. 6299–6308. https://doi.org/10.1109/CVPR.2017.502
8. D.Tran, H.Wang, L.Torresani, J.Ray, Y.LeCun, and M.Paluri, "A closer look at spatiotemporal convolutions for action recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2018, pp. 6450–6459. https://doi.org/10.1109/CVPR.2018.00675

9. J.Huang, W.Zhou, H.Li, and W.Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 9, pp. 2822–2832, 2018. https://doi.org/10.1109/TCSVT.2018.2870740
10. N.C.Camgoz, O.Koller, S.Hadfield, and R.Bowden, "Multi-channel transformers for multi-articulatory sign language translation," in Euro- pean Conf. Computer Vision, 2020, pp. 301–319. https://doi.org/10.1007/978-3-030-66823-5_18
11. M.Boha´cˇek and M.Hru˚z, "Sign pose-based transformer for word- level sign language recognition," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision, 2022, pp. 182–191. https://doi.org/10.1109/WACVW54805.2022.00024
12. Y.Du, P. Xie, M.Wang, X.Hu, Z.Zhao, and J.Liu, "Full transformer network with masking future for word-level sign language recognition," Neurocomputing, vol. 500, pp. 115–123, 2022. https://doi.org/10.1016/j.neucom.2022.05.051
13. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.
14. M.Sandler,A.Howard,M.Zhu,A.Zhmoginov, and L.C.Chen,"MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2018, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
15. S.Hochreiter and J.Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735
16. D.Bahdanau, K.Cho, and Y.Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
17. C.Gu¨ndu¨z and H.Polat, "Turkish sign language recognition based on multistream data fusion," Turkish J. Electr. Eng. Comput. Sci., vol. 29, no. 3, pp. 1171–1186, 2021. https://doi.org/10.3906/elk-2005-156
18. K.K.Podder et al., "Signer-independent Arabic sign language recog- nition system using deep learning model," Sensors, vol. 23, no. 16, p. 7156, 2023. https://doi.org/10.3390/s23167156
19. N.Naz, H.Sajid, S.Ali, O.Hasan, and M.K.Ehsan, "MIPA-ResGCN: A multi-input part attention enhanced residual graph convolutional framework for sign language recognition," Comput. Electr. Eng., vol. 112, p. 109009, 2023. https://doi.org/10.1016/j.compeleceng.2023.109009
20. D.Laines, M.Gonzalez-Mendoza, G.Ochoa-Ruiz, and G.Bejarano, "Isolated sign language recognition based on tree structure skeleton images," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recog- nition, 2023, pp. 276–284. https://doi.org/10.1109/CVPRW59228.2023.00033
21. D.Li,C.Rodriguez, X.Yu,and H.Li,"Word-level deep sign language recognition from video: A new large-scale dataset and methods com- parison," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision, 2020, pp. 1459–1469.
22. A.Akdag and O.K.Baykan, "Enhancing signer-independent recognition of isolated sign language through advanced deep learning techniques and feature fusion," Electronics, vol. 13, no. 7, p. 1188, 2024. https://doi.org/10.3390/electronics13071188
23. D.Kumari and R.S.Anand, "Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism," Electronics, vol. 13, no. 7, p. 1229, 2024. https://doi.org/10.3390/electronics13071229
24. G.Hugar, R.M.Kagalkar, and A.Das, "Comparative study of hybrid deep learning models for Kannada sign language recognition," Int. J. Comput. Intell. Syst., vol. 18, p. 191, 2025. https://doi.org/10.1007/s44196-025-00922-4
25. R.Rastgoo, K.Kiani, and S.Escalera, "Hand sign language recognition using multi-view hand skeleton," Expert Systems with Applications, vol. 213, 2023.