HYBRID CNN-LSTM WITH SOFT ATTENTION FOR REAL-TIME ASL GESTURE RECOGNITION USING MEDIAPIPE LANDMARKS

Akhilesh Rawat[1], Siddharth Chordia[2], Saksham Jain[3], Khush Mangla[4]

[1,2,3,4]Bennett University, Greater Noida, UP, India

ABSTRACT

This paper introduces a real-time American Sign Language (ASL) recognition system using a hybrid deep learning model that integrates 1D Convolutional Neural Networks(Conv1D), Bidirectional Long Short-Term Memory (BiLSTM) layers, and a custom soft attention mechanism. The model processes 60-frame sequences of 3D pose and hand landmarks (258 features per frame) extracted using Mediapipe, allowing the system to learn both spatial and temporal gesture patterns without relying on raw image input. A custom dataset comprising 6,800 gesture sequences across 17 classes—including 16 ASL signs and an idle state—was curated using spatial data augmentation techniques such as mirroring and rotational transformations to ensure robustness across diverse users, lighting, and angles. The model achieved a high testing accuracy of 95.88%, with 96.87% precision and 95.88% recall, and significantly outperformed GRU, RNN, and CNN-RNN baselines in comparative evaluations. Despite its architectural complexity, the system performs efficiently on low-power hardware such as a MacBook Air M2, enabling real-time inference suitable for integration into live video conferencing platforms. The proposed solution offers a scalable, lightweight, and accessible approach to gesture-based communication, advancing the practicality of deep learning in inclusive assistive technologies. Index Terms—LSTM, GRU, RNN, ASL Gesture, BiLSTM, Mediapipe.

KEYWORDS

ASL Gesture Recognition, Hybrid Deep Learning, BiLSTM with Attention, Mediapipe Landmarks, Real-Time Inference.

1. INTRODUCTION

According to the World Health Organisation (WHO), approximately 5% of the world's population experiences hearing disabilities [1]. Sign Language Recognition (SLR) has become an essential tool for fostering inclusive communication by enabling seamless interaction between hearing-impaired individuals and the broader community [2]–[4]. With the rise of virtual meetings and digital communication, the need for real-time SLR systems has become more pressing than ever. Efficient SLR tools help overcome language barriers and promote accessibility across various platforms. Despite progress, many current systems still struggle with continuous gesture recognition, especially under varying lighting conditions, user expressions, and camera angles [5], [6]. Manual interpretation of sign language is not scalable, particularly in digital environments where live translation is essential. This gap presents a compelling opportunity for Artificial Intelligence (AI), particularly deep learning (DL), to provide robust, automated solutions [2]–[4], [7]. DL-based methods—including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and attention mechanisms—have shown promising results in recognizing both static and dynamic signs. However, many of these models

Bibhu Dash et al: ICFCSAI-2025
pp. 247-258, 2026. IJCI – 2026 DOI:10.5121/ijci.2026.150126

are limited to alphabet-level classification or require controlled environments. Recent developments have shown the feasibility of lightweight and real-time systems for SLR using tools like Mediapipe and optimised deep learning architectures [4], [7]. Nonetheless, challenges persist when it comes to deploying these models on low-power, consumer-grade hardware.

Our contributions include designing a hybrid deep learning model that combines Conv1D, BiLSTM, and a custom soft attention mechanism to effectively capture spatio-temporal patterns in ASL gesture sequences. The proposed system utilises 60-frame sequences of 3D hand and pose landmarks comprising 258 features per frame, enabling robust and lightweight input processing. In addition, a custom dataset of 6,800 gesture sequences across 17 gesture classes, including an idle state, was created using mirroring and rotational transformations to ensure diversity in lighting conditions, orientation, and user characteristics. The model demonstrates strong real-time inference capability and generalisation performance, achieving 95.88% accuracy, 96.87% precision, and 95.88% recall, significantly outperforming GRU, RNN, and CNN–RNN baseline architectures.

The motivation behind this work arose from the need to help mute individuals participate more effectively in live meetings and video calls. The objective was to develop a system capable of converting sign language into speech or text in real time, allowing users to express ideas, communicate naturally, and engage seamlessly in digital environments. This work represents a step toward making online communication more inclusive and accessible.

The remainder of the paper is organised as follows. Section II reviews related work in the field of sign language recognition. Section III presents the proposed methodology. Section IV describes the dataset used in this study, while Section V outlines the experimental setup. Section VI presents the results and analysis, followed by visual outputs in Section VII. Section VIII discusses the findings, and Section IX concludes the study.

2. LITERATURE REVIEW

Recent advances in Sign Language Recognition (SLR) have leveraged deep learning and vision-based approaches to improve recognition accuracy and real-time applicability across various sign languages. Baihan et al. [2] presented an improved SLR model based on CNNSa-LSTM, combining CNN, self-attention, and LSTM to effectively capture spatial and temporal features. Using VGG16 and optical flow, along with a hybrid optimizer (HOA and PFA), the model achieved an impressive accuracy of 98.7%. Sundar and Bagyammal [3] proposed a vision-based method for detecting static and dynamic ASL alphabets using Mediapipe hand landmarks and LSTM, achieving 99% accuracy on a custom dataset of 26 alphabets. Mittal et al. [5] introduced a revised LSTM model using Leap Motion for continuous Indian Sign Language (ISL) recognition, achieving 72.3% accuracy for sentence-level recognition and 89.5% accuracy for isolated signs by segmenting gestures into sub-units. García-Gil et al. [4] suggested a lightweight and flexible real-time method for Mexican Sign Language (MSL) detection using six angular features between distal phalanges and the palm, achieving 99% accuracy and F1 score without complex image processing. Guo et al. [8] introduced a hierarchical LSTM (H-LSTM) encoder–decoder model for continuous Sign Language Translation (SLT), using 3D CNNs and viseme-based representations with temporal attention to address gesture variation and word alignment, demonstrating robust performance on both seen and unseen data. Sánchez-Vicinaiz et al. [9] proposed a CNN-based classifier using Mediapipe for static MSL dactylological signs, achieving 84.57% accuracy, 83.33% sensitivity, and 99.17% specificity, making it suitable for real-time deployment on low-power devices. Borges-Galindo et al. [1] developed an RNN-based system for texting MSL using hand and facial expression tracking, achieving 93% accuracy in offline mode and higher accuracy during online operation using 600 video samples.

Barbhuiya et al. [10] employed modified AlexNet and VGG16 architectures combined with a multiclass SVM for ASL alphabet and number classification, achieving 99.82% accuracy with a model efficient enough to operate on CPU-based systems. Lee et al. [11] developed an ASL learning application incorporating real-time recognition using Leap Motion and an LSTM–RNN with a k-NN classifier, achieving 99.44% average accuracy and 91.82% accuracy in 5-fold cross-validation on 2,600 samples. Bhadouria et al. [7] implemented a real-time recognition system using Mediapipe for hand landmark detection and LSTM networks for temporal modelling, enhancing responsiveness compared to image-only systems. Rakhmadi et al. [12] addressed sign language recognition across ASL, ArSL, and BSL using CNN and RNN models, focusing on challenges related to real-time processing, limited datasets, and model generalisation. Renjith et al. [13] introduced a hybrid CNN–RNN model for ISL recognition to handle non-standardised gestures, achieving 98.2% accuracy across 36 ISL signs. Caraka et al. [14] proposed an LSTM model using skeleton-based data extracted via Mediapipe for BISINDO, attaining 92.85% validation accuracy and outperforming earlier LSTM-based approaches. Selvaraju et al. [15] developed a real-time ISL recognition system integrating CNN, YOLOv5, and Hidden Markov Models (HMM) for gesture-to-text conversion during video conferencing via WebRTC, enhancing accessibility. Renjith and Manazhy [16] conducted a comparative analysis between CNN and RNN approaches for ISL recognition, highlighting CNNs' spatial feature extraction strength and RNNs' capability to model temporal dependencies, thereby supporting hybrid CNN–RNN architectures. Vyavahare et al. [6] proposed an LSTM-based framework for dynamic ISL gesture recognition from video, achieving 96% training accuracy and 87% testing accuracy, although further work was required for sentence-level translation. Alabduallah et al. [17] introduced the ISLRHP-HMOADL model, integrating ResNeXt101, VGG19, and Vision Transformers (ViT) with BiGRU for classification and hybrid CS-IGWO optimisation, attaining 99.57% accuracy on an ASL dataset. Venugopalan and Reghunadhan [18] developed a deep convolutional LSTM model for recognising emergency ISL gestures from deaf COVID-19 patients, achieving 83.36% accuracy on their dataset and up to 99.34 ± 0.66% on benchmark datasets. Snehalatha et al. [19] proposed a real-time recognition system using an LSTM model with transfer learning from VGG-16 and ResNet, successfully handling gesture detection in varied environments. Lastly, Kothadiya et al. [20] introduced DeepSign, a deep learning model combining LSTM and GRU layers trained on the IISL2020 dataset, achieving 97% accuracy on 11 ISL signs and offering a practical solution to communication barriers for hearing- and speech-impaired individuals.

3. METHODOLOGY

The proposed model processes sequential landmark data extracted using Mediapipe, consisting of body and hand keypoints obtained from 60-frame gesture sequences. Each frame is represented by a 258-dimensional feature vector comprising 132 features from 33 pose landmarks, including x, y, z, and visibility values, and 126 features from 21 hand landmarks per hand, represented by x, y, and z coordinates. Accordingly, the input sequence is defined as:

$$X = [x_1, x_2, \ldots, x\_T], \text{ where } x\_t \in \mathbb{R}^d \qquad (1)$$

Where, T = 60 represents the number of frames, and d = 258 denotes the feature dimension per frame.

To capture both spatial and temporal dependencies, the proposed architecture integrates three key components: a one-dimensional Convolutional Neural Network (Conv1D), stacked Bidirectional Long Short-Term Memory (BiLSTM) layers, and a soft attention mechanism. Initially, a Conv1D layer with 256 filters is employed to extract local motion patterns from the input sequence. The convolutional operation is expressed as:

$$h^{(1)} = \mathrm{ReLU}(Wc * X + bc) \qquad (2)$$

where, Wc and bc represent the convolutional kernel and bias, respectively, and $*$ denotes the one-dimensional convolution operation. Batch Normalization is subsequently applied to stabilise the learning process, followed by a Dropout layer with a rate of 0.3 to mitigate over fitting.

The extracted convolutional features are then passed through two stacked BiLSTM layers with 128 and 64 units, respectively. These layers model long-range temporal dependencies by processing the sequence in both forward and backward directions. The BiLSTM output at each time step is defined as:

$$\mathrm{H} = [\overrightarrow{h}_t ; \overleftarrow{h}_t], \forall\, \mathrm{t} \in \{1, \dots, \mathrm{T}\} \qquad (3)$$

2, which applies a Multiply layer and Lambda for weighted context aggregation. To selectively focus on important frames, a soft attention mechanism is applied. Attention scores are computed as:

$$et = \tanh(Waht + ba) \qquad (4)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{K=1}^{T} \exp(e_t)} \qquad (5)$$

where Wa and ba are trainable parameters, and αt represents the normalized attention weight. The attention-refined context vector c is obtained by:

$$c = \sum_{t=1}^{T} \alpha_t h_t \qquad (6)$$

This context vector is passed through a Dense layer with 128 ReLU units, followed by dropout(0.4), and then projected to the output space via a softmax classifier:

$$y = \mathrm{Softmax}(Woc + bo) \qquad (7)$$

where Wo and bo denote the output layer's weights and bias, and y denotes the probability distribution over the 17 output classes (16 ASL gestures + 1 Idle State). The model is compiled with the Adam optimizer and categorical cross-entropy loss:

```
model.compile(optimizer='adam',
loss='categorical_crossentropy',
metrics=['accuracy'])
```

The attention mechanism used is illustrated in Figure 2, which consists of a Dense layer, Flatten, Softmax, Repeat Vector, and Permute layers, followed by element-wise multiplication to generate the final context representation. The entire architecture is summarized in Figure 1.

The proposed model's logic is formally summarized in Algorithm 1. The algorithm demonstrates the step-by-step processing pipeline from input to final gesture classification. Algorithm 1 outlines the hybrid deep learning architecture integrating 1D Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM) networks, and an attention mechanism for sequence-based classification tasks. Initially, the input sequence is processed through a 1D CNN layer to extract local features. These features are then fed into stacked BiLSTM layers, which capture temporal dependencies in both forward and backward directions. An attention mechanism is applied over the BiLSTM outputs to assign dynamic importance weights to each time step, enabling the model to focus on the most relevant parts of the sequence. The weighted context vector is passed through a dense layer followed by a softmax activation to

Require: Input sequence $X = [x_1, x_2, \ldots, x_T]$

$h^{(1)} \leftarrow \text{ReLU}(W_c * X + b_c)$ *// 1D Convolution*

Apply BatchNormalization and Dropout(0.3)

$H \leftarrow \text{BiLSTM}_{128,64}(h^{(1)})$ *// Two stacked BiLSTM layers*

for $t = 1$ to T **do**

  $e_t \leftarrow \tanh(W_a h_t + b_a)$ *// Compute attention scores*

end for

$\alpha_t \leftarrow \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}$ *// Normalize weights*

$c \leftarrow \sum_{t=1}^{T} \alpha_t h_t$ *// Generate context vector*

$y \leftarrow \text{Softmax}(W_o c + b_o)$ *// Classification*

return Predicted gesture label $\arg\max(y)$

Algorithm1: Hybrid CNN-BiLSTM with attention for Gesture Recognition

Produce the final output. This architecture is particularly effective for tasks involving long sequences and temporal dynamics, such as gesture recognition or biomedical signal classification.

In conclusion, the model leverages spatial filtering via CNNs, temporal memory from BiLSTMs, and dynamic weighting through attention to effectively recognize ASL gestures from sequence data. Its modular design ensures interpretability and adaptability across real-time gesture-based applications.

4. Dataset Description

The dataset used in this study was specifically curated to support real-time American Sign Language (ASL) gesture recognition. Mediapipe's hand and pose modules were employed to extract three-dimensional landmarks, capturing a total of 17 gesture classes, including 16 static ASL signs and one idle state. Each gesture was recorded as a 60-frame sequence, where each frame consisted of 258 features, including 132 values derived from 33 pose landmarks representing x, y, z, and visibility parameters, and 126 values obtained from 21 hand landmarks per hand, represented by x, y, and z coordinates.

A total of 850 original samples, corresponding to 50 samples per gesture class, were collected from multiple individuals under varying conditions such as lighting intensity, gesture execution speed, and camera angle. This approach ensured both intra-class diversity and inter-user variability within the dataset. To improve model generalisation and better simulate real-world variability, spatial data augmentation techniques were applied. Initially, horizontal mirroring was performed on all samples to accommodate both left-handed and right-handed signing, thereby increasing the dataset size to 1,700 sequences.

Subsequently, rotational transformations were applied along the Z-axis, including clockwise and counterclockwise rotations as well as vertical tilts to the left and right, in order to simulate different viewpoints and camera orientations. Each original and mirrored sequence generated four rotated variants, resulting in a final dataset comprising 6,800 gesture sequences, calculated as 850 × 2 × 4. This comprehensive and augmented dataset enables robust training of gesture recognition models and supports reliable performance across diverse user conditions and input environments.

4.1. Performance Measures

The performance of the proposed model was evaluated using standard classification metrics: accuracy, precision, and recall.

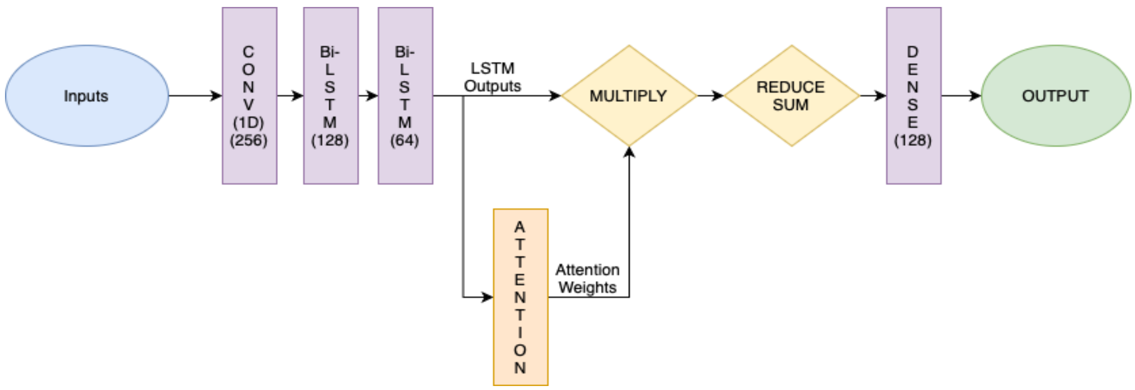

Fig. 1:Attention mechanism within the proposed model

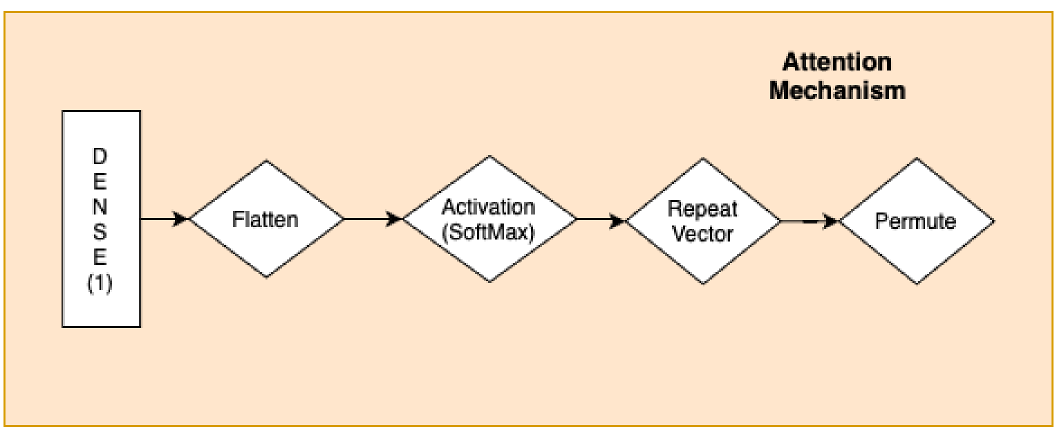

Fig. 2:Proposed model for Sign Language Recognition

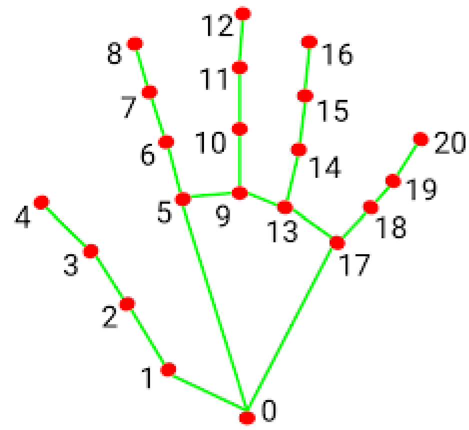

Fig. 3: Hand landmark topology extracted using Mediapipe as spatial features.

These metrics are defined as follows:

$$precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

Where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives. These metrics provide a balanced view of the model's performance, especially in scenarios involving class imbalance.

5. EXPERIMENTAL SETUP

All model training and evaluation were conducted on a MacBook Air M2. The system is equipped with an 8-core Apple M2 processor consisting of four performance cores and four efficiency cores, an integrated 8-core M2 GPU, and 8 GB of unified memory. In addition, the device includes a 16-core Apple Neural Engine. Despite its lightweight and consumer-grade configuration, the system demonstrated sufficient computational capability to support real-time gesture recognition, enabling efficient training and evaluation of deep learning models without the need for GPU-accelerated cloud-based environments.

6. RESULT AND ANALYSIS

The proposed hybrid model incorporating Conv1D, BiLSTM, and a soft attention mechanism, outperforms both the GRU + Dense and RNN-only models across all key performance metrics. RNNs are generally limited in their ability to handle long-range dependencies, and while GRUs offer some improvements through memory gating, they still fall short in effectively capturing spatial features. In our architecture, the Conv1D layer extracts local motion patterns from landmark sequences, while the BiLSTM layers model long term temporal dependencies. The addition of a soft attention mechanism further enhances performance by allowing the model to dynamically focus on the most relevant frames within each gesture sequence, leading to improved. To further evaluate the model's performance, a confusion matrix was generated based on a subset of test samples, as shown in Figure 4.

The matrix illustrates strong diagonal dominance, indicating that most classes were correctly classified. Notably, classes such as when, how, sorry, and thank you achieved perfect or near-perfect classification accuracy. To further evaluate the model's performance, a confusion matrix was generated based on a subset of test samples, as shown in Figure 4. The matrix illustrates strong diagonal dominance, indicating that most classes were correctly classified. Notably, classes such as when, how, sorry, and thank you achieved perfect or near-perfect classification accuracy. Minor misclassifications occurred in similar or visually overlapping signs, such as "drink" and "eat," where a few instances were confused due to overlapping hand shapes or pose transitions. Nevertheless, the matrix demonstrates the model's high level of discriminative ability across all 17 gesture classes, including the IDLE STATE, with overall strong performance even under diverse test conditions.

The results presented in the comparative bar chart demon strate that the proposed CNN + BiLSTM + Attention model consistently outperforms other architectures across all three key metrics: Accuracy, Precision, and Recall. Specifically, it achieves 95.88% accuracy, 96.87% precision, and 95.88% recall, indicating a strong balance between correctly identifying true gestures and minimizing false predictions. In comparison, the CNN+RNN model ranks second across

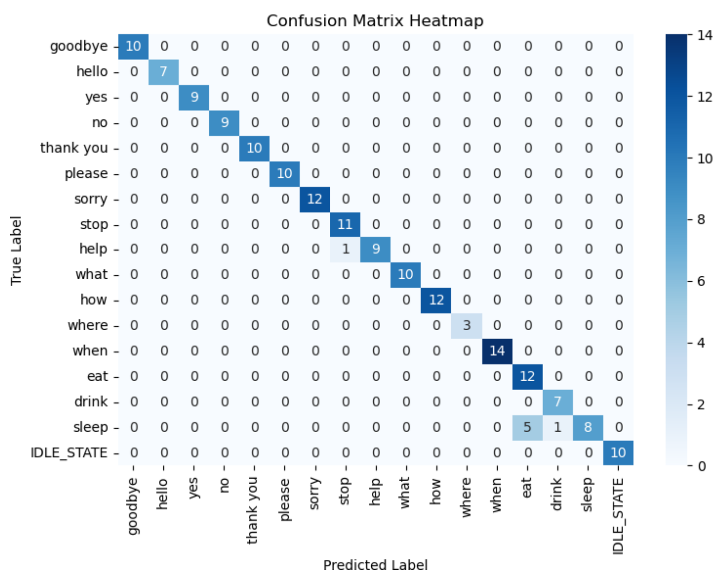

Fig. 4: Confusion matrix showing class-wise prediction performance over a small portion of the test set

most metrics, while Bi-GRU and RNN-only models perform the lowest, particularly in recall and accuracy. When considering the average of all three metrics, the proposed model achieves an average performance of 96.21%, which is significantly higher than CNN+RNN (94.68%), GRU+Dense (93.83%), RNN-only (91.40%), and Bi-GRU (91.13%). This indicates that the proposed hybrid architecture is not only more accurate but also more reliable and consistent across different aspects of performance. The integration of Conv1D for spatial pattern extraction, BiLSTM for temporal modeling, and a soft attention mechanism for dynamic feature weighting proves to be a highly effective combination for real-time sign language recognition.

Overall, the proposed model demonstrates superior generalization and robustness, making it well-suited for deployment in real-world assistive communication systems.

TABLE I: Comparative Evaluation of Model Architecture using Accuracy, Precision, and Recall Metrics

Model Type	Accuracy	Precision	Recall
Proposed	0.9588	0.9687	0.9588
GRU + Dense	0.9365	0.9420	0.9365
RNN Only	0.9121	0.9180	0.9118
CNN + RNN	0.9450	0.9512	0.9441
Bi-GRU	0.9076	0.9198	0.9064

The proposed hybrid model combining CNN, BiLSTM, and an attention mechanism outperforms all other tested architectures, achieving 95.88% accuracy, 96.87% precision, and 95.88% recall [Table 1].

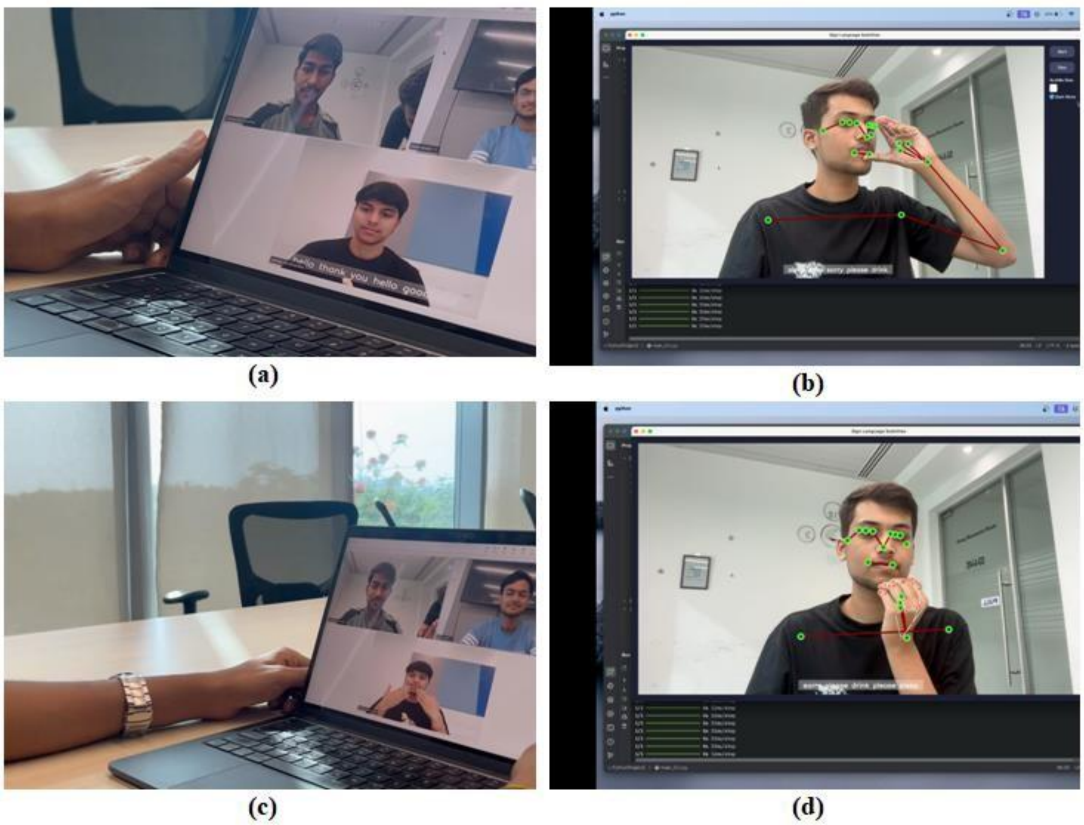

Fig. 5: 5a: Real-time meeting testing interface, 5b: ”Drink” gesture predicted in real time with overlaid Mediapipe hand and pose landmarks, 5c: ”Thank you” predicted in live Microsoft Teams meeting, and 5d: ”Sleep” gesture predicted with corresponding landmarks

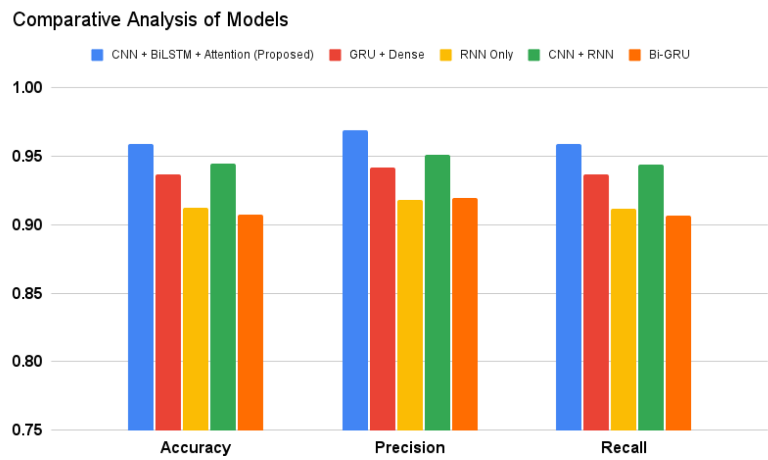

Fig. 6: Comparative Analysis of Models

7. VISUAL REPRESENTATION

This section presents qualitative examples highlighting the real-time capabilities and practical deployment of the proposed ASL gesture recognition system. The selected images demonstrate the model's effectiveness across diverse environments, with Mediapipe hand and pose landmarks overlaid to visualize the spatial features used during prediction. Image 5a depicts the real-time testing interface during a meeting, showcasing the system's responsiveness and integration potential. Image 5b shows the accurate recognition of the "Drink" gesture, with landmark overlays confirming precise spatial understanding. Image 5c highlights the prediction of "Thank you" during a live Microsoft Teams session, demonstrating robustness in streaming environments. Image 5d presents the classification of the "Sleep" gesture, with clear visualisation of the supporting landmarks. Collectively, these examples underscore the model's practical applicability for real-time gesture-based communication systems.

8. DISCUSSION

The proposed hybrid model combining Conv1D, BiLSTM, and a soft attention mechanism demonstrates strong performance in real-time ASL recognition, achieving 95.88% accuracy, 96.87% precision, and 95.88% recall. It effectively captures spatial and temporal features from Mediapipe landmarks, outperforming GRU and RNN baselines, and works well even on consumer-grade hardware like the MacBook Air M2. The augmented dataset and inclusion of an IDLE STATE improved model generalization and real-time usability. Minor misclassifications in similar gestures like "drink" and "eat" suggest potential for refining temporal modeling. Compared to prior works, this model offers greater robustness across users and environments. Its successful integration into video calls highlights its practical value for assistive communication. Future work could focus on continuous sign translation and enhanced gesture tracking using multimodal data or temporal smoothing.

9. CONCLUSION

In this study, we developed an efficient real-time Sign Language Recognition (SLR) system using a hybrid deep learning architecture that combined Conv1D, Bidirectional LSTM, and a soft attention mechanism. We extracted spatial and temporal features from Mediapipe hand and pose landmarks, enabling the model to learn gesture sequences with high precision and reliability. The proposed model achieved strong performance, with 95.88% accuracy, 96.87% precision, and 95.88% recall across 17 ASL gesture classes, including an idle state. We conducted comparative evaluations with baseline models such as GRU, RNN, and CNN+RNN, and the results showed that our model consistently outperformed them in all key metrics. Despite being trained and tested on consumer-grade hardware, the system demonstrated robust real-time performance, making it a practical and accessible solution for communication support in digital environments. Furthermore, we created an augmented dataset to enhance generalization across users and conditions, which helped the model maintain accuracy in varied real-world scenarios. Although the current system focused on isolated gesture recognition, future work may extend it to continuous sign language translation, sentence level recognition, and multi-modal input fusion. Overall, this work provided a promising step toward inclusive and scalable communication technologies for individuals with hearing and speech impairments.

REFERENCES

[1] E. A. Borges-Galindo, N. Morales-Ram´ırez, M. Gonz´alez-Lee, J. R.Garc´ıa-Mart´ınez, M. Nakano-Miyatake, and H. Perez-Meana, “Sign language interpreting system using recursive neural networks,” Applied Sciences, vol. 14, no. 18, p. 8560, 2024.

[2] A. Baihan, A. I. Alutaibi, M. Alshehri, and S. K. Sharma, “Sign language recognition using modified deep learning network and hybrid optimization: a hybrid optimizer (ho) based optimized cnnsa-lstm approach,” Scientific Reports, vol. 14, no. 1, p. 26111, 2024.

[3] B. Sundar and T. Bagyammal, “American sign language recognition for alphabets using mediapipe and lstm,” Procedia Computer Science, vol.215, pp. 642–651, 2022.

[4] G. Garc´ıa-Gil, G. d. C. L´opez-Armas, J. J. S´anchez-Escobar, B. A.Salazar-Torres, and A. N. Rodr´ıguez-V´azquez, “Real-time machine learning for accurate mexican sign language identification: A distalphalanges approach,” Technologies, vol. 12, no. 9, p. 152, 2024.

[5] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, “A modified lstm model for continuous sign language recognition using leap motion,” IEEE Sensors Journal, vol. 19, no. 16, pp. 7056–7063, 2019.

[6] P. Vyavahare, S. Dhawale, P. Takale, V. Koli, B. Kanawade, and S. Khonde, “Detection and interpretation of indian sign language using lstm networks,” J. Intell Syst. Control, vol. 2, no. 3, pp. 132–142, 2023.

[7] A. Bhadouria, P. Bindal, N. Khare, D. Singh, and A. Verma, “Lstm-based recognition of sign language,” in Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing, 2024, pp. 508–514.

[8] D. Guo, W. Zhou, H. Li, and M. Wang, “Hierarchical lstm for sign language translation,” in Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018.

[9] T. J. S´anchez-Vicinaiz, E. Camacho-P´erez, A. A. Castillo-Atoche, M. Cruz-Fernandez, J. R. Garc´ıa-Mart´ınez, and J. Rodr´ıguez-Res´endiz, “Mediapipe frame and convolutional neural networks-based finger spelling detection in mexican sign language,” Technologies, vol. 12, no. 8, p. 124, 2024.

[10] A. A. Barbhuiya, R. K. Karsh, and R. Jain, “Cnn based feature extraction and classification for sign language,” Multimedia Tools and Applications, vol. 80, no. 2, pp. 3051–3069, 2021.

[11] C. K. Lee, K. K. Ng, C.-H. Chen, H. C. Lau, S. Y. Chung, and T. Tsoi, “American sign language recognition and training method with recurrent neural network,” Expert Systems with Applications, vol. 167, p. 114403, 2021.

[12] A. Rakhmadi, A. Yudhana, and S. Sunardi, “A study of worldwide patterns in alphabet sign language recognition using convolutional and recurrent neural networks,” Jurnal Teknik Informatika (Jutif), vol. 6, no. 1, pp. 187–204, 2025.

[13] S. Renjith, R. Manazhy, and M. S. Suresh, “Recognition of sign language using hybrid cnn–rnn model,” in International Conference On Innovative Computing And Communication. Springer, 2024, pp. 17–28.

[14] R. E. Caraka, K. Supardi, R. Kurniawan, Y. Kim, P. U. Gio, B. Yuniarto, F. Z. Mubarok, and B. Pardamean, “Empowering deaf communication: a novel lstm model for recognizing indonesian sign language,” Universal Access in the Information Society, vol. 24, no. 1, pp. 771–783, 2025.

[15] T. Selvaraju, V. Dakshinamurthi, G. Badurudeen, A. Kalyanaraman, A. Kumar, and D. Sampath, “Sign language recognition: Enhancing communication for the hearing and speaking impaired using hybrid model,” in AIP Conference Proceedings, vol. 3279, no. 1. AIP Publishing LLC, 2025, p. 020126.

[16] S. Renjith and R. Manazhy, “Indian sign language recognition: A comparative analysis using cnn and rnn models,” in 2023 International Conference on circuit power and computing technologies (ICCPCT). IEEE, 2023, pp. 1573–1576.

[17] B. Alabduallah, R. Al Dayil, A. Alkharashi, and A. A. Alneil, “Innovative hand pose based sign language recognition using hybrid metaheuristic optimization algorithms with deep learning model for hearing impaired persons,” Scientific Reports, vol. 15, no. 1, p. 9320, 2025.

[18] A. Venugopalan and R. Reghunadhan, “Applying hybrid deep neural network for the recognition of sign language words used by the deaf covid-19 patients,” Arabian Journal for Science and Engineering, vol. 48, no. 2, pp. 1349–1362, 2023.

[19] N. Snehalatha, S. M. Kumar, H. Kavitha, S. Nisha, P. Gowda, and S. K. Prasad, “Sign language detection using action recognition lstm deep learning model,” in 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON). IEEE, 2024, pp. 1–6.

[20] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-Gonz´alez, and J. M. Corchado, “Deepsign: Sign language detection and recognition using deep learning,” Electronics, vol. 11, no. 11, p. 1780, 2022.

AUTHORS

Akhilesh Rawat has completed Ph.D. from Jawaharlal Nehru University, New Delhi, India. He is working as an Assistant Professor, Bennett University, Noida. His research focused on Generalization in Deep Learning and Machine Learning. He has also completed a Master of Science in Computer Science from Banaras Hindu University, Varanasi.

Siddharth Chordia2 is a B.Tech Computer Science and Engineering student with interests in machine learning, data science, and software systems. His work focuses on building practical, data-driven applications and exploring applied artificial intelligence.

Saksham Jain3 is an engineering student at Bennett University, India, with research interests in artificial intelligence, deep learning, and computer vision. His work focuses on applying machine learning techniques to real time systems, with an emphasis on practical and impactful solutions.

Khush Mangla4 is currently pursuing B.Tech in Computer Science and Engineering at Bennett University. His research interests include generative AI, machine learning, and computer vision.