DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp. 717-724, 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Real-Time Sign Language Recognition using MediaPipe and 1D CNN

Mamta Bisht, Kumud Kundu, Saksham Bhardwaj, Gaurav Sharma, Ankit Sharma

Department of Computer Science and Engineering (AIML), Inderprastha Engineering College, Ghaziabad, India

mamtabisht@ipec.org.in, kumud.kundu@ipec.org.in, shakshambhardwaj2003@gmail.com, gaurav.sharma122000@gmail.com, as9650377@gmail.com

Abstract

Sign language recognition is a critical technology that enhances communication for individuals who are deaf or hard of hearing. This study presents a real-time sign language recognition system that combines MediaPipe for efficient feature extraction and a 1D Convolutional Neural Network (CNN) for accurate gesture classification. Our approach focuses on dynamic American Sign Language (ASL) signs and addresses key challenges in the field, including real-time performance, hardware cost, and adaptability. The dataset comprises 15 dynamic ASL signs, recorded under various lighting conditions and backgrounds to ensure robustness. The proposed model achieves 98% accuracy in recognising the 15 dynamic ASL signs. Notably, our system operates in real time with only a standard webcam, making it both affordable and easily deployable.

1. Introduction

Sign language is an essential form of communication for individuals with hearing or speech impairments. Automated recognition of sign language has gained significant traction due to advancements in computer vision and deep learning [1], [2]. Modern systems can now recognise both isolated and continuous gestures using visual cues such as hand posture, body orientation, and facial expressions [3], [4]. However, real-time performance, hardware costs, and adaptability remain key challenges [5]. To address these, we propose a real-time recognition system focused on dynamic American Sign Language (ASL) signs. It uses MediaPipe for efficient feature extraction and a 1D CNN for gesture classification. The key objectives of this work are:

- Develop an efficient sign language recognition system using MediaPipe and a 1D CNN.
- Achieve high recognition accuracy across varied environmental conditions.
- Reduce computational costs and training time.

2. Literature Survey

Sign language recognition research has evolved through various methods, each offering distinct advantages and presenting unique challenges [6]. Early approaches primarily relied on sensor-based devices, such as data gloves, to track hand movements [7]. Although these methods provided precise motion data, they were often intrusive and hindered the natural expression of signs. In contrast, recent advancements in computer vision and deep learning have enabled the development of vision-based sign recognition systems. These approaches eliminate the need for physical sensors and instead utilize video input to capture gestures. Researchers have explored a variety of deep learning architectures, such as Convolutional Neural Networks

(CNNs) and Recurrent Neural Networks (RNNs), to recognise both isolated and continuous signs [8]. Comparative analyses of sign language recognition techniques highlight the effectiveness of deep learning models in capturing complex spatial-temporal patterns inherent in signing. Publicly available datasets such as MS-ASL [1] and BSL-1K [9] have significantly advanced this field by providing large-scale training and evaluation resources. Despite these advances, there remains a pressing need for real-time, low-cost, and robust sign recognition systems that can be easily deployed across diverse environments. Sign language recognition systems can broadly be categorized into sensor-based and vision-based methods. Sensor-based systems, which often utilize gloves or motion capture suits, provide detailed motion tracking but are typically expensive, cumbersome, and unnatural for daily use [10]. On the other hand, vision-based systems apply machine learning techniques to recognize gestures from video frames, offering a more natural and scalable solution.

These vision-based methods commonly extract visual features such as hand shape, motion trajectories, and body posture to classify gestures using neural networks. However, one of the major challenges in this domain is handling continuous signing, where gestures flow seamlessly and exhibit co-articulation. Unlike isolated gestures, continuous sign language requires models to interpret temporal dependencies and temporal variations. To address this, researchers have employed advanced architectures such as recurrent neural networks and temporal convolutional networks, which are well-suited for sequence modelling. Another significant challenge is the diversity of sign languages across different regions. Each sign language has its own unique vocabulary, grammar, and visual-spatial characteristics. Hence, there is a need for adaptable recognition systems that support multiple sign languages, ensuring inclusivity for deaf communities worldwide. While state-of-the-art vision-based systems have achieved high accuracy, they often rely on complex neural networks and high-end computational resources, which limit their deployment in real-time or resource-constrained environments. To overcome these limitations, the present work proposes a lightweight, real-time sign recognition system that uses MediaPipe for efficient feature extraction and a 1D CNN for gesture classification. Related efforts, such as those by [11] and [12] have explored sign language recognition using MediaPipe in conjunction with deep learning models. These studies demonstrated the accuracy and effectiveness of MediaPipe in detecting hand landmarks and joint positions, which were then used for classification tasks. While those works focused on Assamese and American Sign Language, our proposed system specifically targets dynamic American Sign Language gestures, emphasizing real-time performance and adaptability in everyday environments.

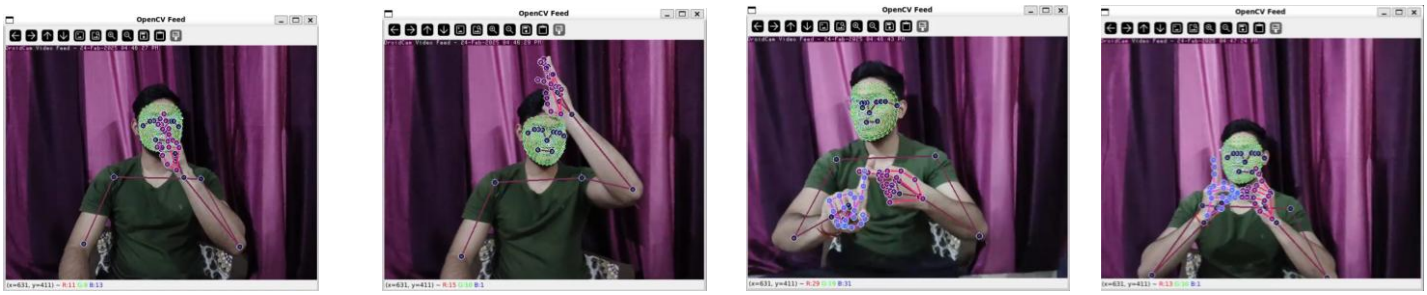
3. Dataset Description

The dataset consists of 15 dynamic ASL signs (e.g., Mother, Father, Hello, Help) captured using a standard webcam, with landmark coordinates, angles, and distances extracted via MediaPipe for robust real-time recognition. Each sign was recorded in various lighting conditions and backgrounds to ensure robustness. Screenshots of each label in the dataset are shown in Figure 1. The dataset includes:

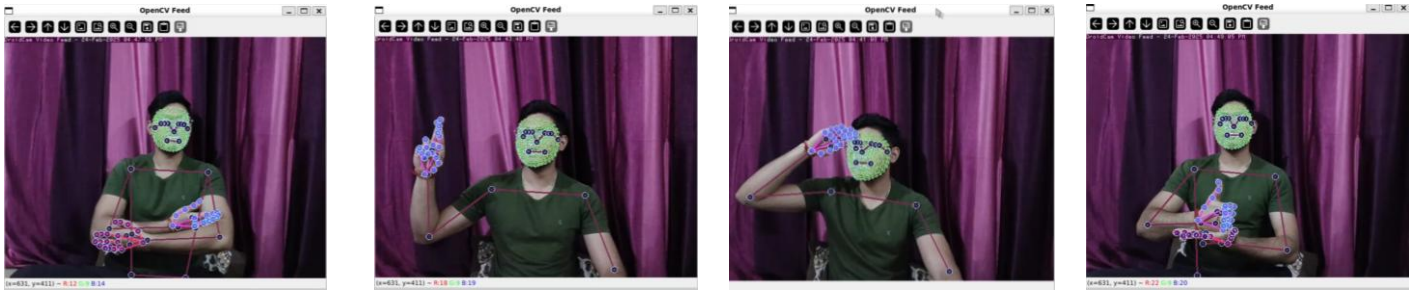
- Landmark coordinates (X, Y, Z) extracted using MediaPipe.

- Angles between key joints.
- Distances between selected landmarks.

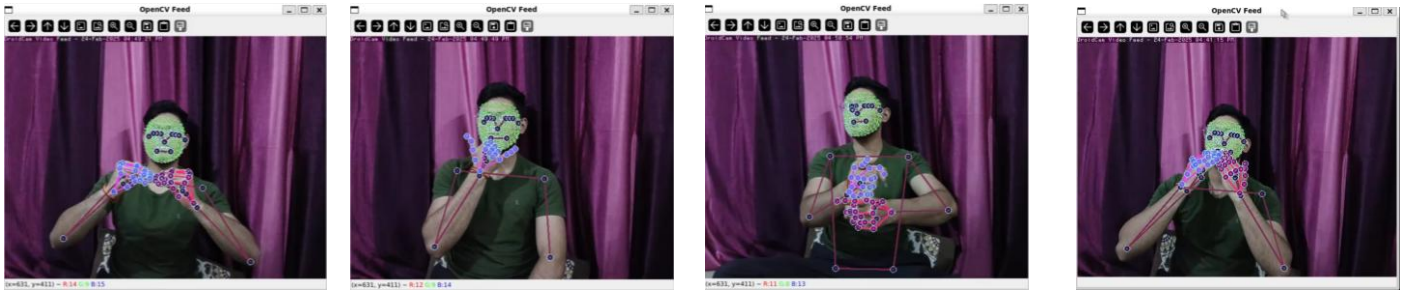
The dataset was split into 80% for training and 20% for testing to evaluate the model's performance.



Mother Father Friend Dating



Baby Uncle Hello Help



More Wrong Right Thank you

Phone Eat You

Fig 1: Screenshots of each label in the dataset are shown.

4. Proposed model

The system follows a **two-phase** approach:

1. **Feature Extraction:** MediaPipe extracts key hand and body landmarks. Fig 2 and Fig 3 show 21 3D landmark points on a hand and 33 3D landmark points on a pose from a single frame respectively.
2. **Sequence Classification:** CNN processes extracted features to classify the sign. The 1D CNN model summary is given in Table 1.

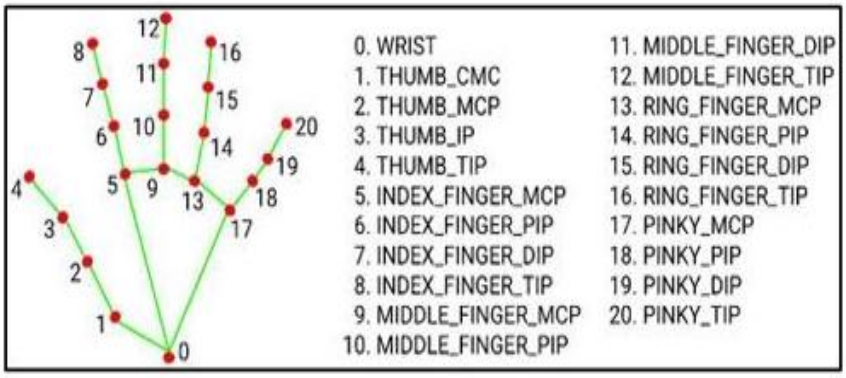


Fig: 2: Hand Landmark

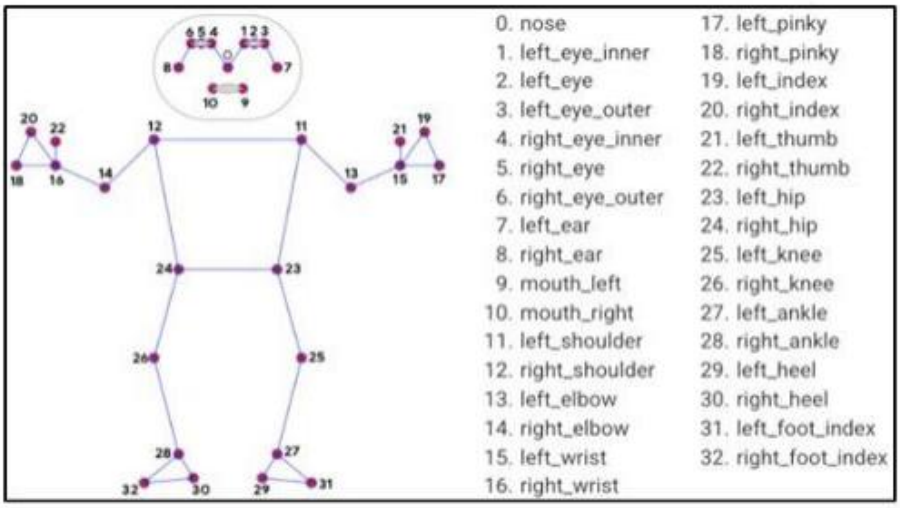


Fig: 3: Pose Landmark

Table 1: Summary of proposed 1D CNN model

Layers (type)	Output shape	Parameters
Conv1D	(None, 28, 64)	319168
Maxpooling1D	(None, 14, 64)	0
Conv1D	(None, 12, 128)	24704
Maxpooling1D	(None, 6, 128)	0

Conv1D	(None, 4, 64)	24640
Maxpooling1D	(None, 2, 64)	0
Flatten	(None ,128)	0
Dense	(None, 64)	8256
dropout	(None, 64)	0
Dense_1	(None, 32)	2080
Dense_2	(None, 15)	495
Total parameters: 1138031 Trainable parameters: 379343 Non- Trainable parameters: 0		

The proposed model combines MediaPipe for feature extraction as shown in Figure 4 and 1D CNN for sequence classification. MediaPipe is used to extract hand and body landmarks as presented in Figure 4, which are then processed by the 1D CNN model to classify the sign. The CNN model consists of three layers with 128, 64, and 64 filters, respectively, followed by dense layers for classification. The model uses Adam optimizer and softmax activation for the final output and was trained for 1500 epochs, achieving 98% accuracy. The proposed model demonstrates high efficiency, with a training time of approximately four minutes and a recognition time of less than one second, enabling rapid deployment and real-time performance.

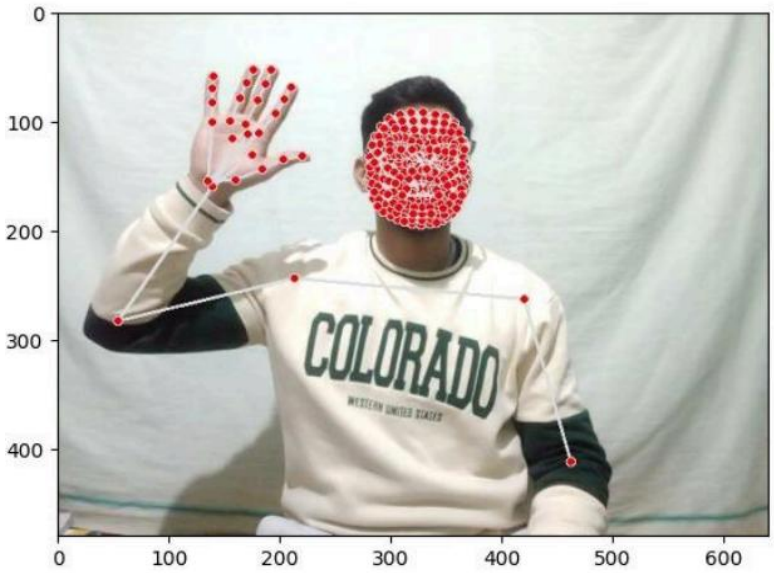
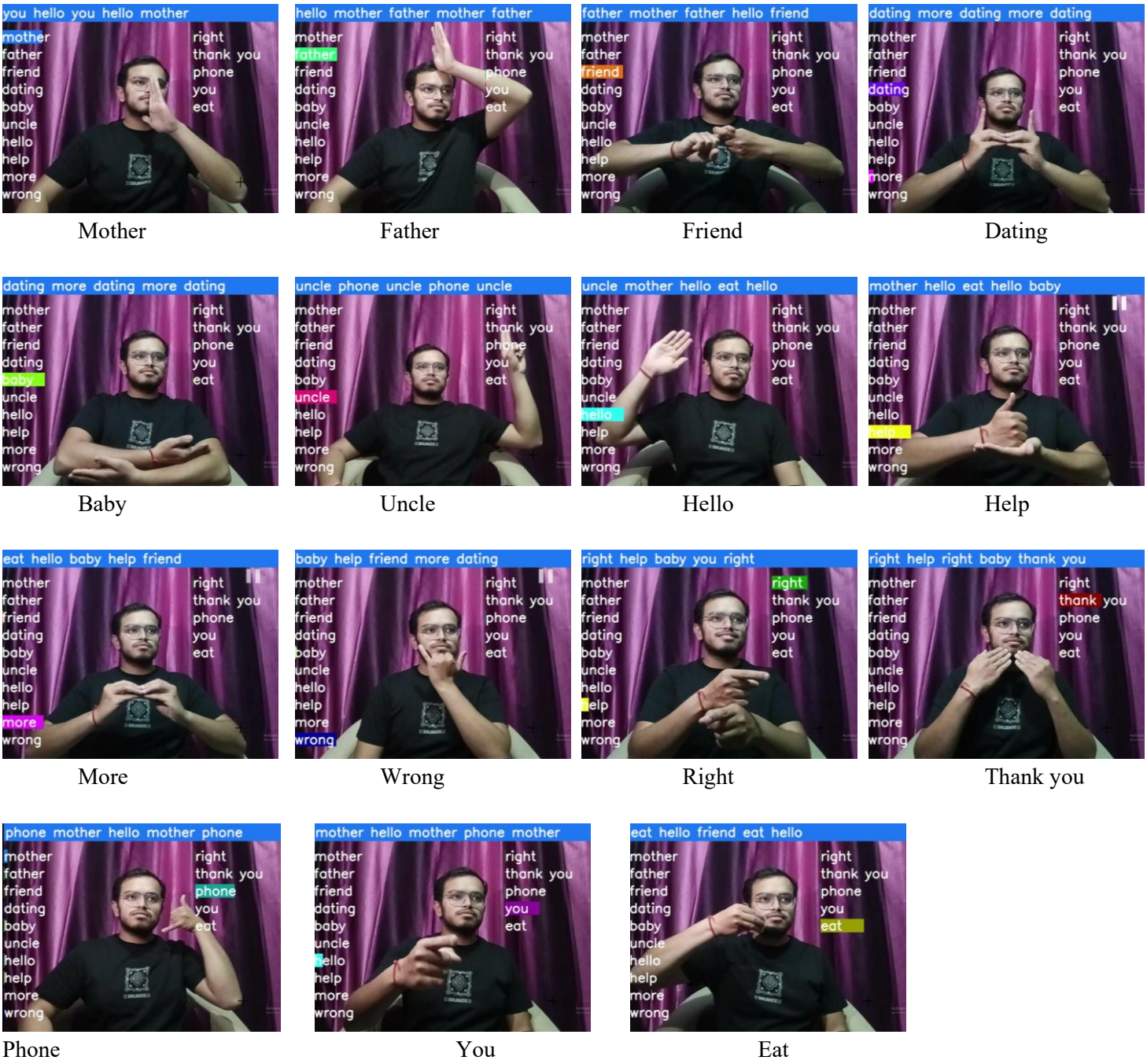


Fig 4: Feature extraction from Mediapipe setup

5. Result

The system was tested using real-time video input and benchmarked against existing methods. For each label, some screenshots of the results are shown in Figure 5. The proposed model consists of three 1D convolutional layers with ReLU activations, followed by max pooling and

dense layers. Dropout is used to prevent overfitting. The model was trained using the Adam optimizer and categorical cross-entropy loss. The Test Accuracy achieve is 98% for 15 dynamic ASL signs.



Mother Father Friend Dating

Baby Uncle Hello Help

More Wrong Right Thank you

Phone You Eat

Fig: 5 - Detecting dynamic ASL gestures.

The confusion matrix is presented in Figure 6. It provides a visual representation of the model's performance across the 15 dynamic ASL signs. Table 2 compares the proposed method with other existing approaches for sign language recognition.

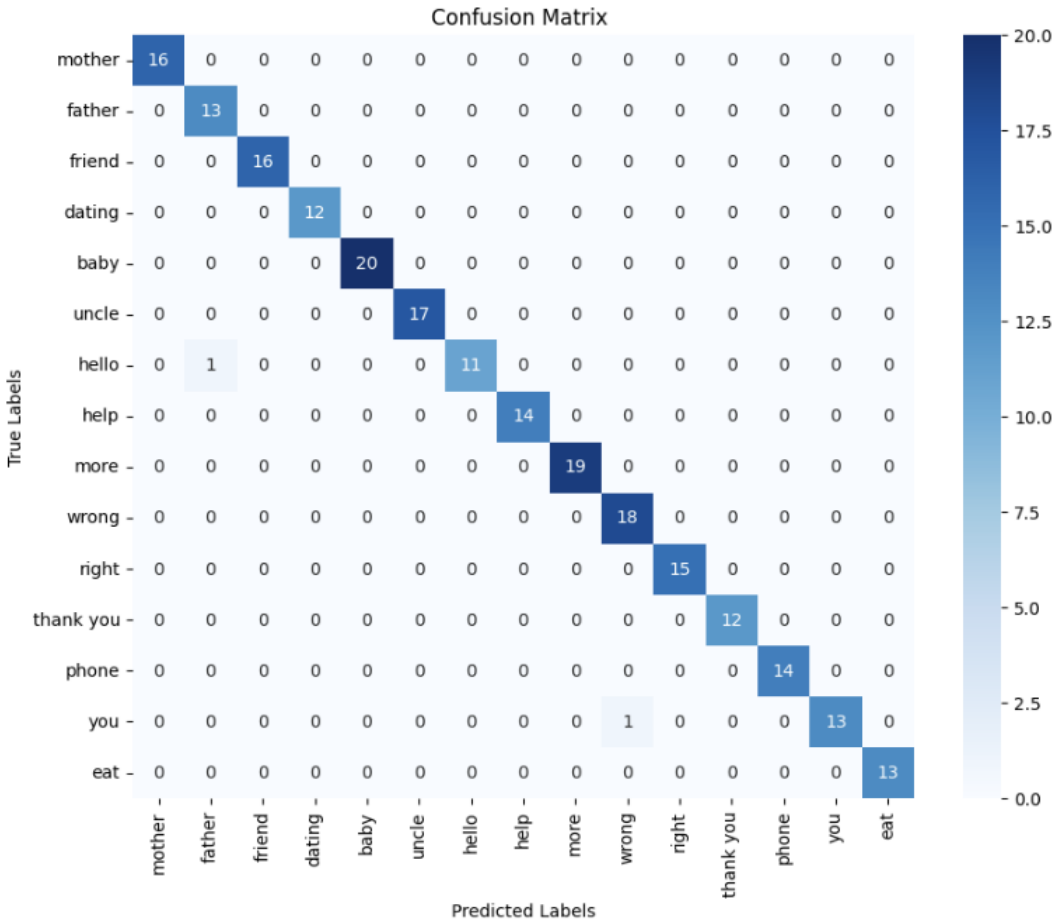


Fig 6: Confusion matrix

Table 2: Comparison with Other Methods

Author	**Method**	**Hardware requirement**	**Real-time capability**	**Accuracy**
Gałka et al., 2016 [7]	Sensor gloves	High (glove-based)	No	85%
Kumar et al., 2017 [13]	HMM	Medium	No	88%
Proposed model	MediaPipe and 1D CNN	Low (Webcam)	Yes	98%

6. Conclusion

The proposed work utilizes MediaPipe to extract hand and body landmarks from video input captured by a standard webcam. These features, including 3D landmark coordinates, angles between key joints, and distances between selected landmarks, are then processed by a 1D CNN model for real-time ASL recognition. This research contributes to the field of sign language recognition by offering a lightweight, accurate, and adaptable solution. The combination of MediaPipe and a 1D CNN proves effective at capturing the spatial-temporal patterns inherent in signing while maintaining low computational requirements. It achieves 98% accuracy using only a standard webcam, making it affordable and easy to deploy. Future

work will focus on expanding the gesture set, supporting continuous signing, and deploying the model on mobile or embedded platforms.

References

[1] Joze, H. R. V., & Koller, O. (2018). Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*.

[2] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Sign language recognition via skeleton-aware multi-model ensemble. *arXiv preprint arXiv:2110.06161*.

[3] Thakar, S., Shah, S., Shah, B., & Nimkar, A. V. (2022, October). Sign language to text conversion in real time using transfer learning. In *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)* (pp. 1-5). IEEE.

[4] Srivastava, S., Gangwar, A., Mishra, R., & Singh, S. (2021, December). Sign language recognition system using TensorFlow object detection API. In *International conference on advanced network technologies and intelligent computing* (pp. 634-646). Cham: Springer International Publishing.

[5] Zhou, W., Zhao, W., Hu, H., Li, Z., & Li, H. (2025). Scaling up multimodal pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[6] Kumar, R., Sinha, A., Bajpai, A., & Singh, S. K. (2023). A comparative analysis of techniques and algorithms for recognising sign language. *arXiv preprint arXiv:2305.13941*.

[7] Gałka, J., Mąsior, M., Zaborski, M., & Barczewska, K. (2016). Inertial motion sensing glove for sign language gesture acquisition and recognition. *IEEE Sensors Journal*, *16*(16), 6310-6316.

[8] Renz, K., Stache, N. C., Albanie, S., & Varol, G. (2021, June). Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2135-2139). IEEE.

[9] Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., & Zisserman, A. (2020, August). BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European conference on computer vision* (pp. 35-53). Cham: Springer International Publishing.

[10] Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2021). Artificial intelligence technologies for sign language. *Sensors*, *21*(17), 5843.

[11] Sundar, B., & Bagyammal, T. (2022). American sign language recognition for alphabets using MediaPipe and LSTM. Procedia Computer Science, 215, 642-651.

[12] Bora, J., Dehingia, S., Boruah, A., Chetia, A. A., & Gogoi, D. (2023). Real-time assamese sign language recognition using mediapipe and deep learning. *Procedia Computer Science*, *218*, 1384-1393.

[13] Kumar, P., Gauba, H., Roy, P. P., & Dogra, D. P. (2017). Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, *86*, 1-8.