International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Special Issue ICCIS-2025, pp 144-149, May-2025

https://doi.org/10.47001/IRJIET/2025.ICCIS-202523

Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

MediaPipe and Deep Learning for Robust Real-Time Hand Gesture Recognition in Sign Language

[1]D Sumathi, [2]Potteti Tejaswini, [3]Sadineni Aasritha, [4]Gadamsetty Deepthika

[1,2,3,4]Department of Computer Science Engineering, Alliance School of Advanced Computing, Bengaluru, India
E-mails: [1]d.sumathiphd@gmail.com, [2]tejaswinipotteti3@gmail.com, [3]aasrithas019@gmail.com, [4]Deepthikag2005@gmail.com

***Abstract* - This extend centers on creating an AI-based framework for real-time sign dialect discovery utilizing computer vision and profound learning methods. The essential objective is to bridge the communication hole between the hard of hearing and hearing communities by precisely recognizing hand motions and changing over them into content or discourse. The examination includes utilizing MediaPipe Hands, OpenCV, and a profound learning show prepared on a dataset of sign dialect signals. Strategies such as convolutional neural systems (CNNs) and repetitive neural systems (RNNs) are utilized to make strides motion acknowledgment precision.**

The MediaPipe Hands system, combined with OpenCV, empowers vigorous real-time hand following and keypoint extraction. Profound learning models, especially CNN-based models, accomplish tall precision in classifying sign dialect motions. The framework performs well in controlled situations but faces challenges with varieties in lighting, foundation clutter, and hand occlusions. Growing the dataset and coordination more complex worldly models (e.g., LSTMs or Transformers) can assist upgrade acknowledgment exactness. Move forward dataset differing qualities by joining more hand shapes, skin tones, and lighting conditions. Execute transient modeling methods (e.g., LSTMs, Transformers) to improve acknowledgment of ceaseless sign dialect.

Keywords: Sign Language Recognition, Real-time Detection, Deep Learning, MediaPipe Hands, OpenCV, CNN, RNN, LSTM, Transformers, Hand Gesture Recognition, Temporal Modeling, Edge Deployment, Multimodal Interpretation

I. INTRODUCTION

The deaf and hard-of-listening to network by and large makes use of signal language, an advanced and expressive approach of communication, that is critical for constructing sturdy interpersonal relationships and facilitating get admission to information, education, and different sides of everyday life. The introduction of technology that near the space among signal language customers and the overall public is essential considering hundreds of thousands of people during the sector use signal language as their number one shape of communication. Here, we introduce a singular studies undertaking called "MediaPipe and Deep Learning for Robust Real-Time Hand Gesture Recognition in Sign Language."

These days, signal language is extensively utilized by the ones who've speech or listening to issues to communicate. Sign language is made of many complicated motions, every of which has a completely unique meaning. If one wishes to talk with a person who has listening to loss, they ought to both be talented with signal language or pay a human translator due to the fact there may be presently no sincere manner to translate the use of computers. Currently, there aren`t many techniques for detecting signal language. Specifically, item detection is employed. Presently, item detection is used to pick out the hands, and then they're transmitted to a signal language version that has already been skilled to carry the preferred message from the user. One such approach which could pick out gadgets and convey an output relying for your skilled version is known as YOLO (You Only See Once).

YOLO will contend with maximum of the work; we simply want to educate the essential model. You won`t want to do whatever on this. Convolution neural networks, which might be able to deep getting to know and direct getting to know from information, could be some other example. In this manner, the community can also additionally obtain the essential information and use it to decide the signal language whilst the digital digicam flow is proven to it.

The World Health Organization (WHO) estimates that five billion humans global be afflicted by a few types of listening to loss, consisting of 430 million who're almost deaf, and that few people are professional in signal language, so having a powerful approach of translating signal language is essential. Thus, it's miles critical to have a straightforward approach for translating signal language. Because it might be tough for one man or woman to translate signal language each time, the process has to additionally be automated. This is due to the fact there aren`t many folks that realize signal language, and when you consider that there are not many translators, mastering it might be pretty expensive. We want to cope with the hassle earlier than we will expand a realistic solution.

II. LITERATURE REVIEW

This demonstrates the usage of ASR, YOLO, and the Google voice-to-textual content API in real-time signal language identity and translation. It will comprise more than one cameras and microphones to file audio and visible information, with a view to then be processed with the aid of using a microprocessor. While a the front digital digicam concentrates on spotting signal language motions, a rear digital digicam will file on any capability threats withinside the surrounding area. By capturing, analyzing, translating, and showing this statistics on a clever glass HUD, deaf people are capable of talk and get hold of real-time remarks approximately their environment[1]. An Effective Deep Learning-Based Real-Time Indian Sign Language (ISL) Detection System: The implementation idea for real-time Indian signal language gesture popularity and type making use of a gadget that integrates CNN, Media Pipe, OpenCV, and different frameworks is proven on this study. It makes use of OpenCV for gesture identity the usage of contour and side detection and Media Pipe for the maximum correct hand detection and tracking. A dataset of ISL gestures is used to educate the deep CNN. The version then optimizes itself in characteristic extraction and type, taking into consideration real-time application. With a few traits that permit text-to-voice speech synthesis [2].

AR glasses for deep learning-primarily based totally signal language recognition: In order to offer device imaginative and prescient processing for signal language images, the item integrates the RNN and YOLO algorithms. On the only hand, YOLO demonstrates high-velocity and high-accuracy item identity in person frames. RNN, on the opposite hand, can procedure sequential statistics, permitting the machine to shop and make use of statistics from in advance frames. This approach can growth the detection`s accuracy and resilience to troubles like item occlusion or movement blur. Images are preprocessed, YOLO is used for item detection, non-most suppression is carried out to the results, RNN is used to label the objects observed primarily based totally on a video sequence, and the final results is displayed on an truth screen [3].

Design of Artificial Neural Network Backpropagation for SIBI Sign Language Recognition: The SIBI alphabet turned into utilized by the authors of this paper to collect high-decision pictures for hand movement detection in signal language the usage of a Kinect sensor; the letters J and Z had been overlooked due to the fact they're dynamic. TensorFlow Lite is used for hand key factor popularity after those pix had been cropped to spotlight the hand area with the reference factor at the bottom of the palm. These salient factors are stored to be used as capabilities in backpropagation schooling of an Artificial Neural Network (ANN). This paper makes use of a dataset of hand key factor coordinates to educate an synthetic neural network (ANN) to discover the gesture [4]. Implementing Security Access Control with Deep Learning-Based American Sign Language Recognition in: Dataset preparation, version training, evaluation, and inferencing are the principle workflow techniques that make up the gadget on this paper. First, an preliminary dataset is created the usage of eighty Kaggle snap shots which have been gotten smaller to 2 hundred via way of means of 2 hundred pixels. Annotation creates XML documents with item coordinates the usage of LabelImg.py. Cutting edge To save you overfitting, YOLOv3, that's brief and powerful for real-time item identification, became paired with statistics augmentation approaches. In order to evaluate accuracy and efficiency, imply common precision is used to estimate version performance. stay circulation detection to assure dependable operation for all great enter types [5].

Convolutional Neural Network-Based Real-Time Gesture Detection: In this research, a CNN structure for gesture identity in video processing is presented, wherein schooling relies upon on the amount and sort of data. Any CNN technique usually begins off evolved with the convolution and max-pooling layers. At the begin of any CNN process, those layers might unavoidably be critical for figuring out low-stage capabilities and decreasing computational load. The gadget uses Maxpooling and Sigmoid activation.

Layers that had been absolutely linked had been critical for classification. It is skilled the usage of gesture-precise and numerical moves for human-gadget interplay the usage of a threshold photo dataset in black and white, which has been streamlined to decrease complexity and boom processing efficiency [6]. Real-time Yolov5 Sign Language Recognition in: Data collection, preprocessing, configuration, and version education are the 4 number one additives of the study, that's primarily based totally on YOLOv5 for gesture recognition. A overall of 2,365 pix of hand gestures in diverse lighting fixtures instances that constitute the Hindi/Marathi alphabets, ASL, and static gestures were gathered. Data changed into augmented via way of means of labeling it with bounding bins and dividing it into education, validation, and checking out sets. Model parameters had been accompanied while enhancing configuration files. It underwent a 12-hour education method that worried two hundred iterations on a Colab Notebook and switch studying at the COCO dataset [7]. Real-time Yolov5 Sign Language Recognition in: Data collection, preprocessing, configuration, and version education are the 4 number one additives of the study, that's primarily based totally on YOLOv5 for gesture recognition. A overall of 2,365 pix of hand gestures in diverse lighting fixtures instances that constitute the Hindi/Marathi alphabets, ASL, and static gestures were gathered. Data changed into augmented via way of means of labeling it with bounding bins and dividing it into education, validation, and checking out sets. Model parameters had been accompanied while enhancing configuration files. It underwent a 12-hour education method that worried two hundred iterations on a Colab Notebook and switch studying at the COCO dataset [8].

Recognition of Chinese Sign Language in Real Time Based on synthetic neural networks, the researchers seize the sEMG indicators of ten sufferers even as they make Chinese signal language motions the usage of the MYO wristband. seventy five education samples and 450 take a look at samples have been recorded for every difficulty at a frequency of 2 hundred Hz sign shooting as a part of the statistics gathering process. Normalization, rectification, and occasional by skip filtering were implemented to uncooked sEMG indicators so as to save you noise. The sliding window approach turned into used for function extraction, generating 376 capabilities according to pattern from time-area capabilities which includes RMS and MAV in addition to pre-processed indicators. The class turned into achieved usage of three-layer synthetic neural community that turned into educated the usage of full-batch gradient descent and the cross-entropy fee feature to are expecting signal language actions with amazing accuracy [9].

Using SSD-Mobile Net to stumble on Indian Sign Language in actual time: This paintings specializes in item detection, which incorporates the SSD-Mobile Net version`s unique detection and localization of letters from A to Z. In order to understand items of various sizes, it's going to first are expecting from several function maps at exclusive resolutions after extracting capabilities from photographs the usage of the Mobile Net structure withinside the SSD-Mobile Net framework. It allows the speedy and unique detection of numerous items in a unmarried picture without the want for bounding container proposals that have been required in preceding approaches. Its open-supply TensorFlow Object Detection API framework simplifies version building, training, and deployment. Tensor Board visualization is supported, and the pre-skilled fashions from the TensorFlow Model Zoo are reused [10].

III. METHODOLOGY

In our research paper we will be using Media Pipe, an inbuilt library for the sign language detection. The key thing that we are doing is:

1. Hand Landmark Detection using MediaPipe: Our approach uses MediaPipe, a dependable and powerful framework for hand monitoring in actual time. It makes it feasible to discover 21 critical landmarks at the human hand, along with the palm and finger joints. These landmarks offer the baseline records for similarly investigation.
2. Landmark Coordinate Extraction: We extract the three-D coordinates (x, y, z) for every of the 21 spots after figuring out the hand landmarks. The hand`s course and posture are captured via way of means of the spatial facts furnished via way of means of those coordinates.
3. Computation of Euclidean Distances: We calculate the Euclidean distances among selected pairs of factors for you to extract huge records from the landmarks that have been recovered. The traditional formulation is used on this calculation: $\mathbf{d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}}$. These distances useful resource in figuring out the overall orientation and posture of the hand in addition to hand configurations like open or closed fingers.
4. Gesture/Sign Classification: We assign the contemporary hand movement to a corresponding signal language image with the aid of using combining the landmark coordinates with the calculated distances. These traits permit us to infer the hand`s direction, position, and shape—all of which might be important markers of numerous indications.
5. Real-Time Detection and Display: The complete pipeline is installation to function in actual time. The detected signal is dynamically proven at the display screen subsequent to the user`s hand after a stay video circulate is processed body via way of means of body. This ensures a dynamic and responsive signal language popularity machine suitable for sensible uses.

IV. FORMULAS AND GRAPHS

Additional Feature Calculations:

1. Angle between Landmarks: We use the cosine rule to calculate the perspective created through 3 landmarks if you want to realise the orientation of arms or hand posture.

   Formula:

   $$\boldsymbol{\theta = \cos^{-1} \left(((x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1) + (z_2 - z_1)(z_3 - z_1)) / (\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} * \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2 + (z_3 - z_1)^2}) \right)}$$

2. Coordinate Normalization: The 2D landmark coordinates are normalized when it comes to the hand`s bounding container in an effort to make certain scale invariance.

   Formula:

x_norm = (x - x_min) / (x_max - x_min)
y_norm = (y - y_min) / (y_max - y_min)

This ensures constant interpretation at numerous hand sizes and digital digicam distances.

Figure 1: MediaPipe Hand Landmark Model

Index	Joint Name
0	Wrist
1–4	Thumb (MCP, IP, TIP)
5–8	Index Finger (MCP, PIP, DIP, TIP)
9–12	Middle Finger (MCP, PIP, DIP, TIP)
13–16	Ring Finger (MCP, PIP, DIP, TIP)
17–20	Pinky Finger (MCP, PIP, DIP, TIP)

Here Joint attitude and distance estimations are primarily based totally on those features.

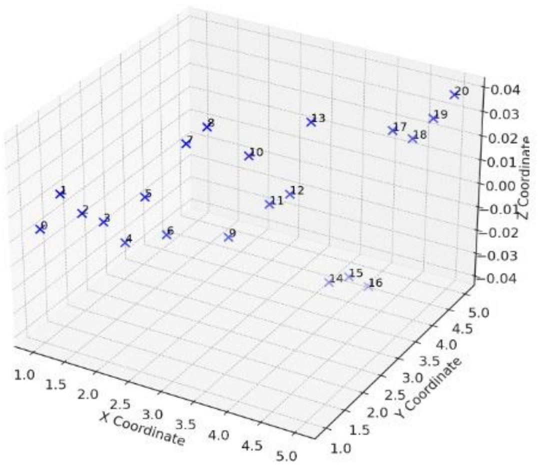


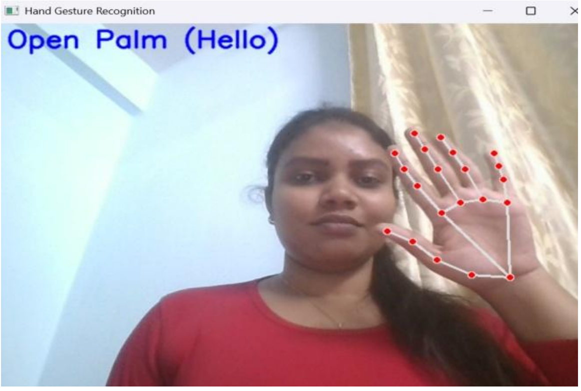
Figure 2: 3D Plot of MediaPipe Hand Landmarks

The spatial association of the 21 hand landmarks that MediaPipe usually detects in a hand gesture reputation device is depicted on this 3-d scatter map. Each blue dot, that's categorized with an index starting from zero to 20, represents one of the hand`s critical points.

1. X and Y Axes: A distribution of hand landmarks at some stage in the breadth and peak of the hand is simulated with the aid of using the X and Y axes, that have values among 1 and 5.
2. Z Axis: There is little intensity change, as indicated with the aid of using the fairly smaller values, which variety approximately between -0.04 and 0.04. Hands usually seem in a 2D aircraft with faint 3-D intensity records in real-international digital digicam settings.
3. Interpretation: In basical planar layout, the landmarks may be used to categorize gestures in keeping with their relative positions and shapes. For example: A "Peace Sign" might suggest a more separation among the landmarks of the index and center fingers.

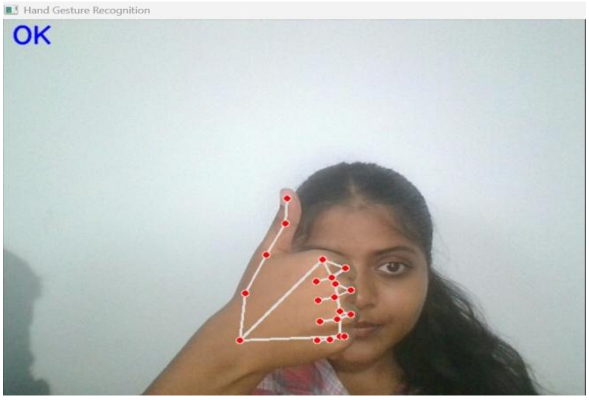
V. RESULTS

Currently there are only 7 signs namely yes, no, Hello, OK, Not OK, I love you, and peace, other than these it shows undefined gesture.



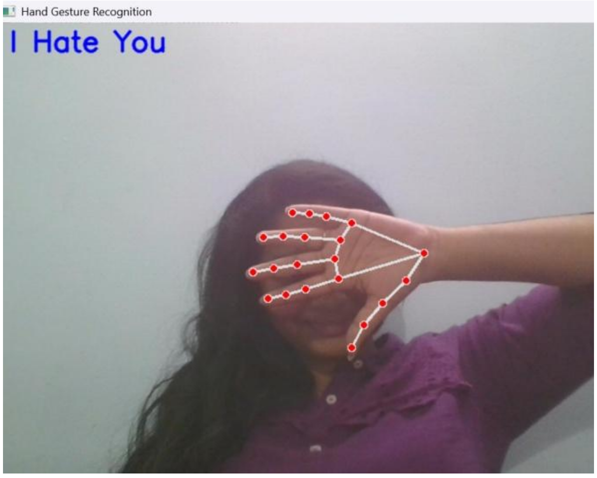
Open Palm (Hello)

Figure 1: When the flat hand shown its represents “Open Palm (Hello)”



OK

Figure 2: When the Thumb is Up It represents “Thumbs Up (OK)”



I Hate You

Figure 3: When the flat hand shown opposite direction it represents " I Hate You"

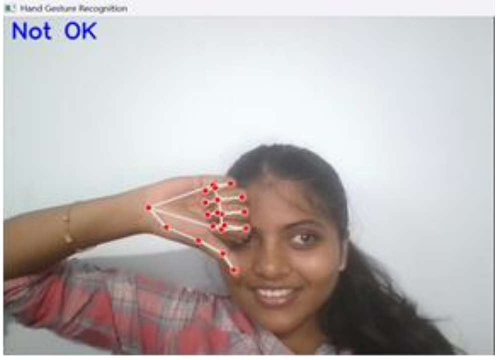


Figure 4: When the Thumb is down It represents "Thumbs Down (Not OK)"

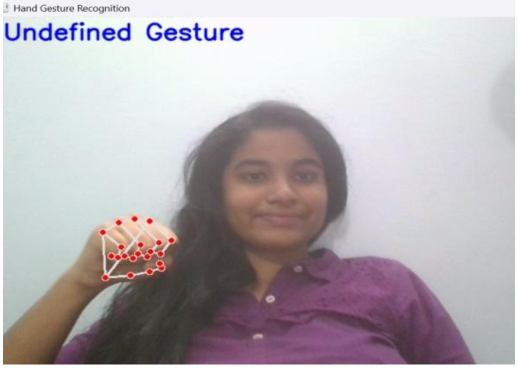


Figure 5: Undefined Gesture

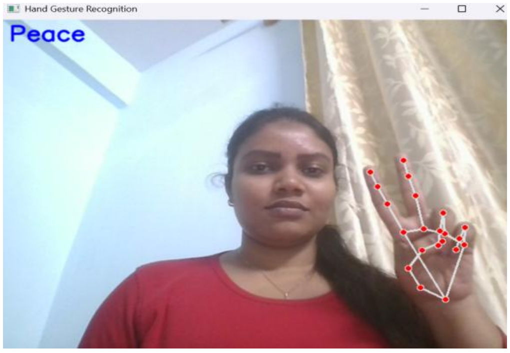


Figure 6: When the 2 fingers, the middle and the index finger are extended it means "peace"

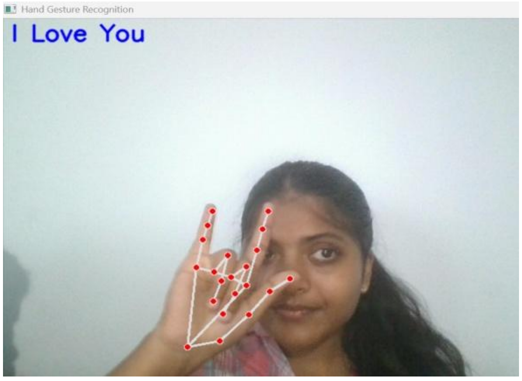


Figure 7: When the Index finger, Thumb finger, and Little Finger extended it Shows "I Love You"

VI. MODULES

Module 1: Focuses on hand landmark detection using MediaPipe, a framework designed for real-time hand tracking. It begins by initializing the MediaPipe Hands module, which sets up the environment for detecting and tracking hand landmarks. The system then captures video input through the webcam to continuously monitor hand movements. Using MediaPipe, the module identifies 21 key landmarks or joints on the hand and retrieves their coordinates in 3D space, including the x, y, and z values. This information is crucial for understanding hand gestures and movements in various applications such as gesture recognition and human-computer interaction.

Module 2: Involves the extraction of landmarks and calculation of distances between specific hand joints. Once MediaPipe detects the hand, the module retrieves the 3D coordinates (x, y, z) of the 21 key landmarks. These coordinates are then used to compute Euclidean distances between relevant joint pairs—such as between the thumb and index finger—to analyze hand gesture signals and spatial relationships. The distance is calculated using the formula:

$$\text{Distance} = \sqrt{}((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2)$$

Module 3: Focuses on signature classification using the calculated distances and angles between joints. It implements a gesture recognition logic that interprets the spatial arrangement of landmarks to classify specific hand signs, such as a thumbs-up or a peace sign. The system uses predefined gesture mappings to associate these detected patterns with their respective meanings, enabling accurate recognition of static hand gestures.

Module 4: Integrates all previous modules to enable real-time sign language detection and visualization. It captures live video through the webcam and overlays recognized gestures directly onto the video feed, placing the identified sign next to the user's hand. This provides immediate and intuitive feedback for users, making the system suitable for real-time gesture-based communication.

REFERENCES

[1] Cai, R., Janaka, N., Kim, H., Chen, Y., Zhao, S., Huang, Y., & Hsu, D. (2025, April). AiGet: Transforming Everyday Moments into Hidden Knowledge Discovery with AI Assistance on Smart Glasses. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-26).

[2] Roy, S., Maiti, A. K., Dutta, B., Basak, G. K., & Ghosh, K. (2025). Digital pedagogy in Indian sign

language: requirement analysis and assistive technology based application. Universal Access in the Information Society, 1-20.

[3] Zhao, S., & Ozaki, T. (2025). Virtual Touchpad for AR Glasses Based on Gesture Recognition. Journal of Information Processing, 33, 128-138.

[4] Yue, H., Wei, Y., Yuan, H., & Li, H. (2025). Revitalizing urban industrial heritage: Enhancing public trust in government through smart city development and open big data analysis using artificial neural network (ANN) modeling. Cities, 156, 105538.

[5] Antari, N. W. A., Riastini, P. N., & Wirabrata, D. G. F. (2025). American Sign Language for Science Terminology in Fourth Grade Elementary School. Jurnal Ilmiah Sekolah Dasar, 9(1), 186-193.

[6] Abdoos, M., Rashidi, H., Esmaeili, P., Yousefi, H., & Jahangir, M. H. (2025). Forecasting solar energy generation in the Mediterranean region up to 2030–2050 using convolutional neural networks (CNN). Cleaner Energy Systems, 10, 100167.

[7] Ma, Qian, et al. "Intelligent Hand-Gesture Recognition Based on Programmable Topological Metasurfaces." Advanced Functional Materials 35.1 (2025): 2411667.

[8] Chen, Y., & Wu, Y. (2025). Detection of Welding Defects Tracked by YOLOv4 Algorithm. Applied Sciences (2076-3417), 15(4).

[9] Liu, Z. L. (2025). Artificial neural networks. In Artificial Intelligence for Engineers: Basics and Implementations (pp. 175-190). Cham: Springer Nature Switzerland.

[10] Smart glasses integrate ASR, YOLO, and voice-to-text APIs to detect and translate sign language in real time, providing deaf users with immediate feedback via a HUD display.

[11] A CNN-based system using MediaPipe and OpenCV enables real-time Indian Sign Language recognition with speech synthesis.

Citation of this Article:

D Sumathi, Potteti Tejaswini, Sadineni Aasritha, & Gadamsetty Deepthika. (2025). MediaPipe and Deep Learning for Robust Real-Time Hand Gesture Recognition in Sign Language. In proceeding of Second International Conference on Computing and Intelligent Systems (ICCIS-2025), published in *IRJIET*, Volume 9, Special Issue ICCIS-2025, pp 144-149. Article DOI https://doi.org/10.47001/IRJIET/2025.ICCIS-202523

****