Maharshi Karve Stree Shikshan Samstha's
Cummins College of Engineering for Women, Pune
(An autonomous Institute affiliated to Savitribai Phule Pune University)
Department of Computer Engineering

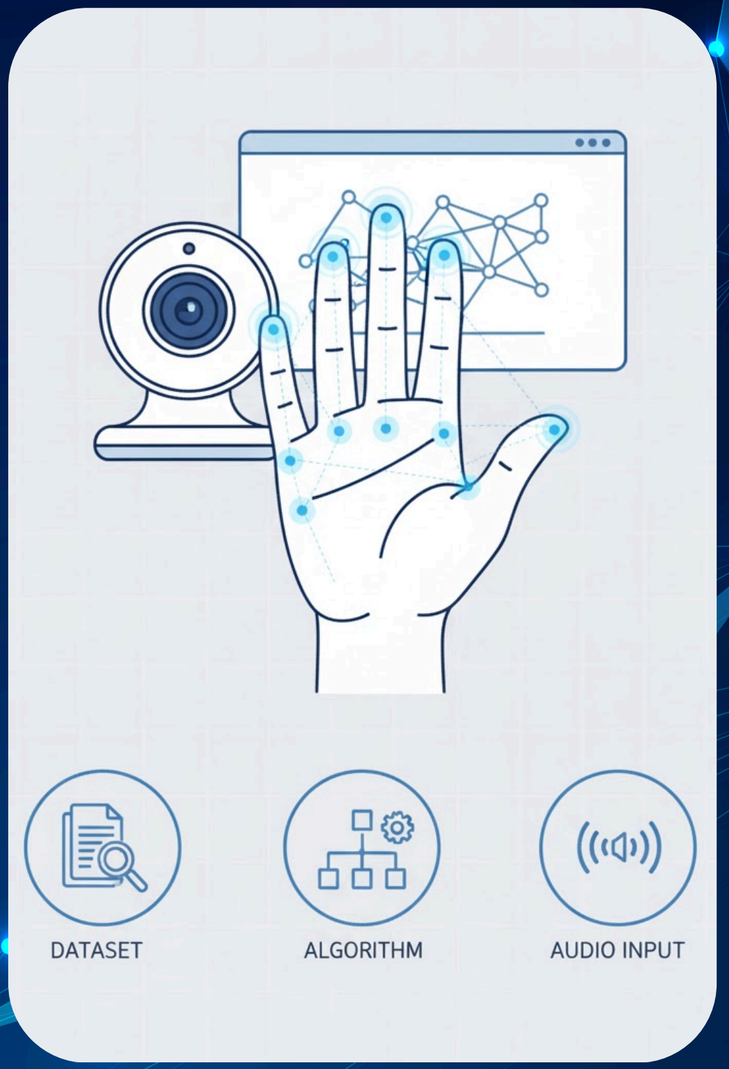
HAND GESTURE RECOGNITION USING MEDIAPIPE AND ML

Our Team:
UCE2023437 Prashika Lonkar
UCE2023438 Sharayu Madage
UCE2023442 Samruddhi Mane

Ty Comp ,div :-A ,batch :- A3

INTRODUCTION

- Hand gesture recognition is important in human–computer interaction, accessibility for speech-impaired users, gaming, and robotics.
- Vision-based methods (camera) are cheaper and simpler compared to glove-based sensors.
- Existing deep-learning models need large datasets and GPUs.
- MediaPipe provides lightweight landmark detection suitable for real-time applications.
- Objective: build a fast, accurate A–Z sign recognition system with audio output for learning and communication support.

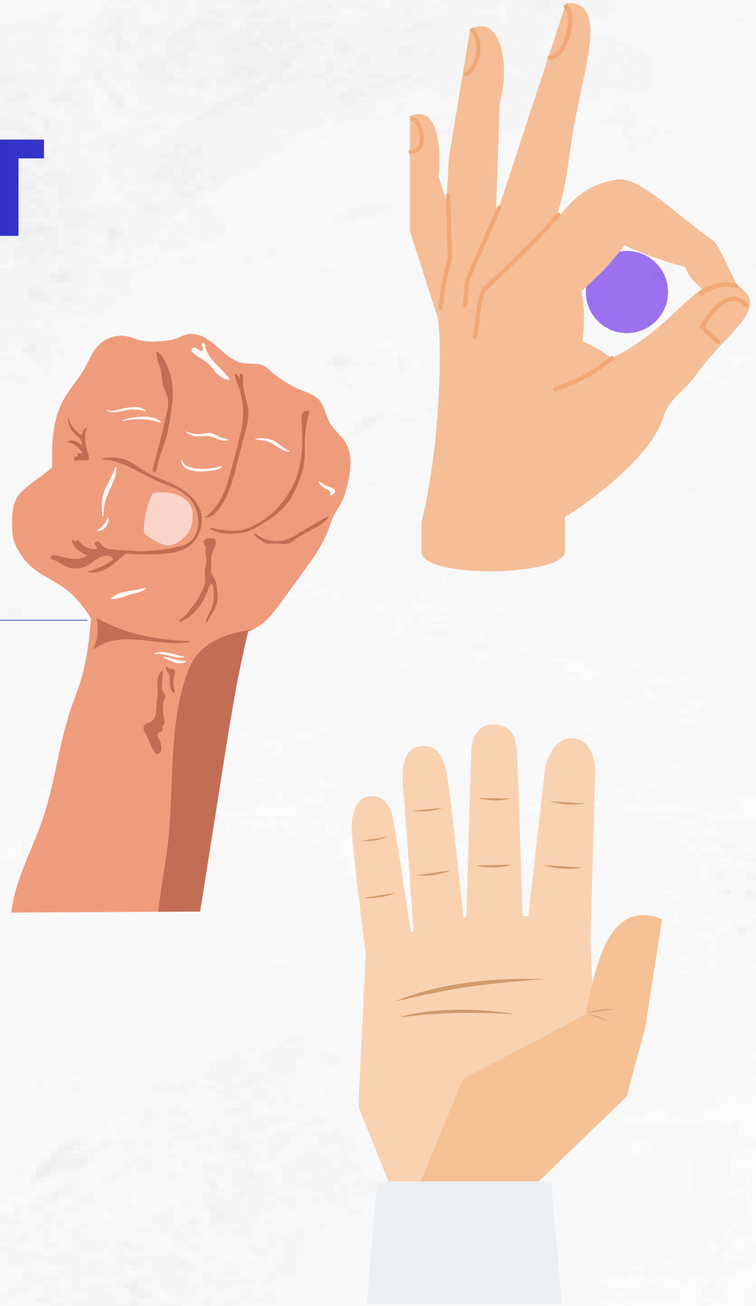


PROBLEM STATEMENT

Traditional input devices like keyboards and buttons are not always suitable for:

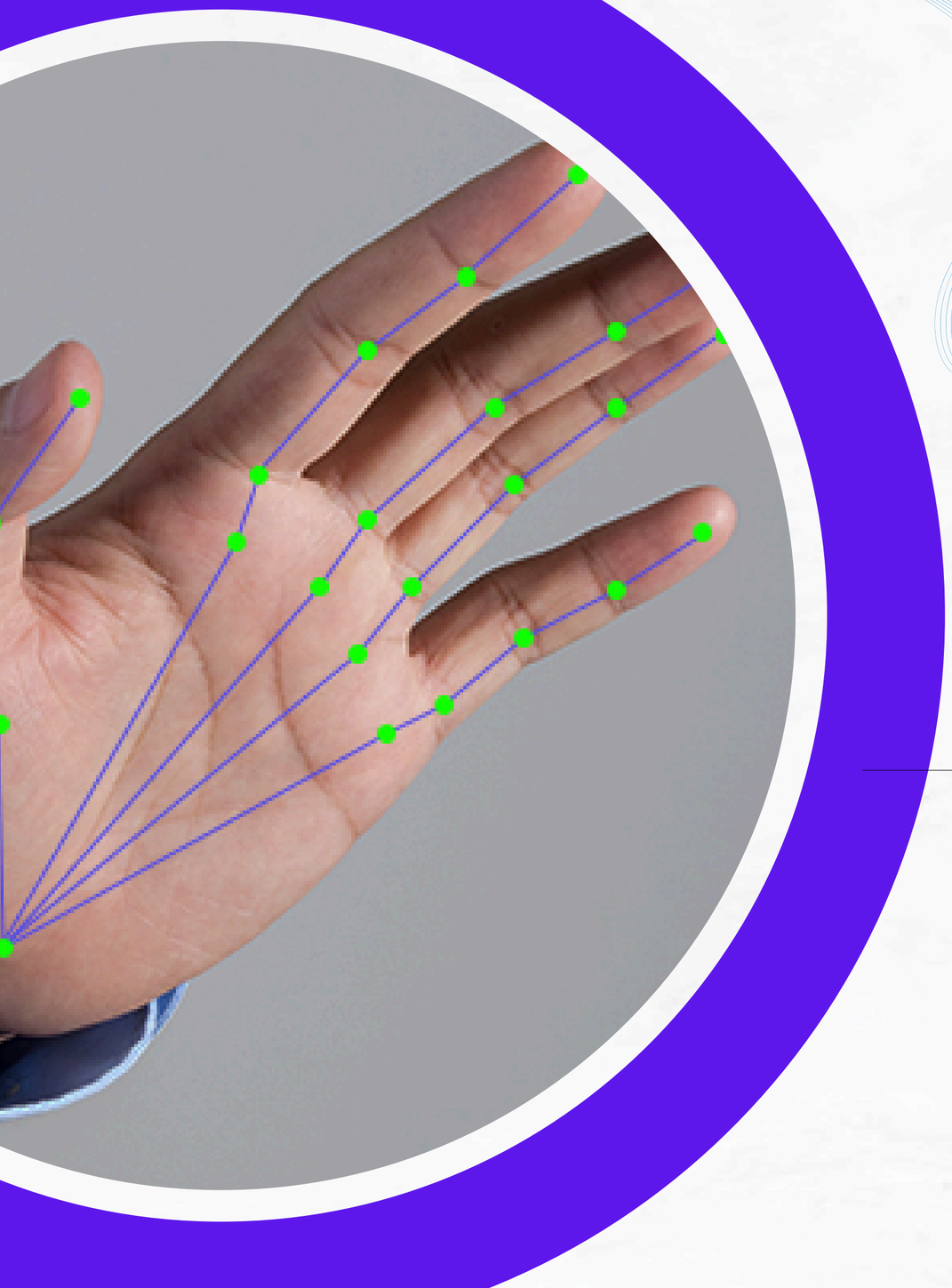
- Physically disabled users
- Sterile environments (labs, hospitals)
- IoT device interaction
- AR/VR gesture-based systems

Thus, a hands-free, real-time gesture recognition system is needed to improve accessibility, speed, and user experience



ABSTRACT

- THIS PROJECT PRESENTS A REAL-TIME HAND GESTURE RECOGNITION SYSTEM FOR PREDICTING ENGLISH ALPHABETS (A–Z).
- MEDIAPIPE HANDS IS USED TO EXTRACT 21 KEY HAND LANDMARKS FROM VIDEO FRAMES.
- A RANDOM FOREST MACHINE LEARNING MODEL CLASSIFIES GESTURES BASED ON LANDMARK FEATURES.
- A FLASK WEB INTERFACE PROVIDES REAL-TIME DETECTION, PREDICTION, AND AUDIO OUTPUT FOR EACH RECOGNIZED GESTURE.
- THE SYSTEM SHOWS HIGH ACCURACY (96%) AND WORKS EFFICIENTLY IN REAL-TIME CONDITIONS, HELPING ACCESSIBILITY AND LEARNING APPLICATIONS.



OBJECTIVES

Implement a real-time hand gesture recognition system using MediaPipe landmarks.

Predict English alphabets (A–Z) with high accuracy.

Use Random Forest ML model for classification.

Integrate audio output for each detected alphabet for accessibility.

Ensure system is fast, scalable, and user-friendly via Flask web interface.

SYSTEM ARCHITECTURE

01

Input:
Webcam Video Capture
- **Real-time stream processed frame by frame**

02

Hand Landmark Detection:
MediaPipe Hands
- **Extracts 21 keypoints from hand image**

03

Feature Extraction & Classification:
42-Dimensional Landmark Feature
- **Normalized co-ordinates Random Forest Classifier**
- **Predicts gesture from features**

04

Output:
Audio Assistance
- **Text-to-Speech for each predicted gesture**

Web Display and Stored Logs

AIML TECHNIQUES USED

1 MediaPipe Hands

- Extracts 21 landmarks
- Lightweight, fast, works with webcam

2 Random Forest Classifier

- 200 trees, depth 20
- Strong accuracy & generalization

3 Med
Flask Framework
Real-time browser interface

4 Libraries

- OpenCV,NumPy,Flask,pyttsx3
- Scikit-learn

COMPARISON OF ALGORITHM WHY USED!

Feature	Random Forest Classifier (Used in Project)	Decision Tree Classifier
Accuracy	95-98%	82-86%
Model Type	Ensemble of multiple trees	Single tree
Overfiiting Risk	Very Low (Trees averaged)	Very High
Stabilliy	Stable — minor changes don't affect output	Unstable — small dataset changes cause new structure
Real-Time Performancce	Fast — supports 30+ FPS	Fast but less accurate
Generalization	Excellent even with noisy hand data	Poor generalization
Dataset Requirement Use Case Fit	Works well with medium data (2600 images)	Tends to memorize data

METHODOLOGY

Dataset Collection:

- 2600 images captured manually (100 per alphabet A–Z).
- Different lighting, backgrounds, orientations for better generalization.

Feature Extraction:

- MediaPipe extracts 21 hand landmarks → converted to 42 normalized features (x, y).
- Normalization removes scale and position differences.

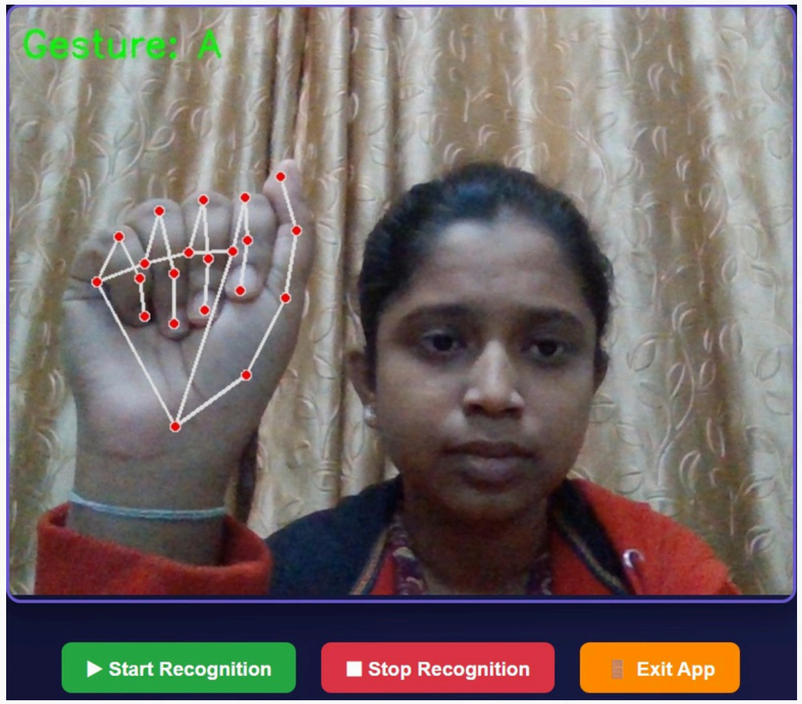
Model Training:

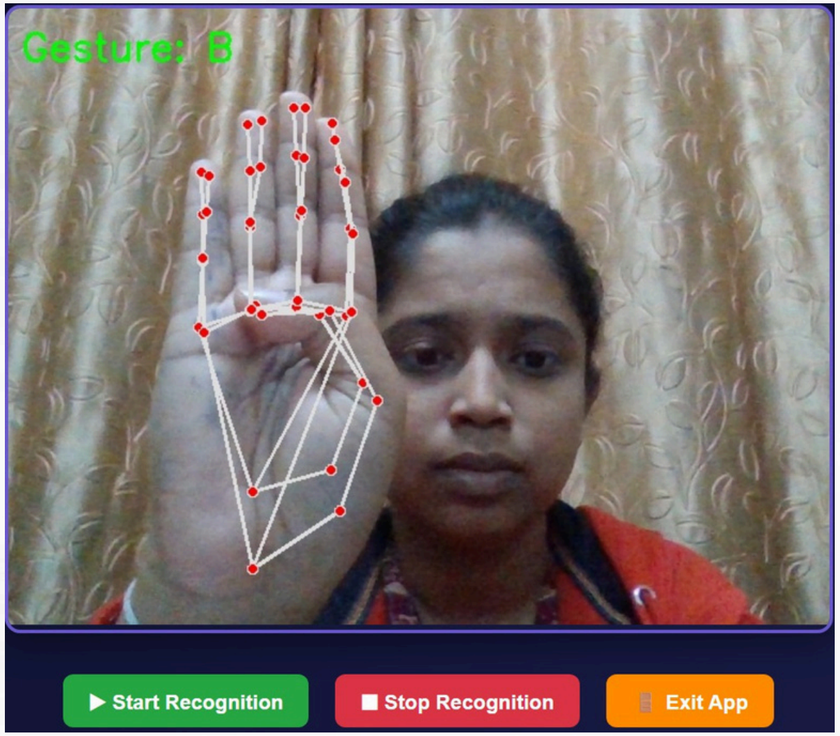
- Random Forest Classifier with 200 trees and depth 20.
- Train–test split: 80–20.
- Achieved 95–98% accuracy.

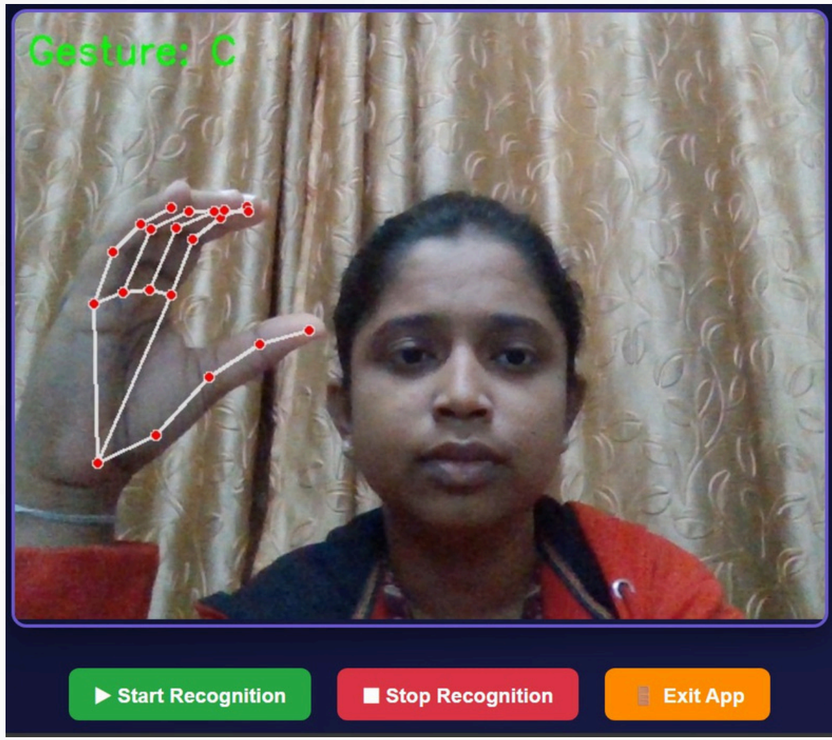
Real-Time System:

- Flask web app + OpenCV webcam stream.
- Every frame is processed: detection → landmarks → prediction → audio.
- Cooldown avoids repeated audio for the same gesture

RESULTS & EVALUATION

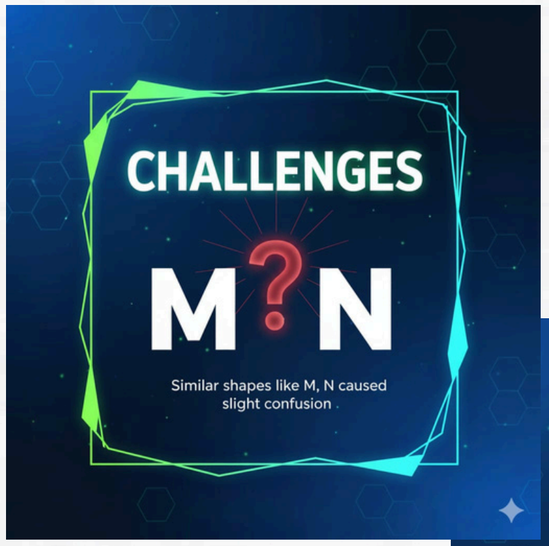


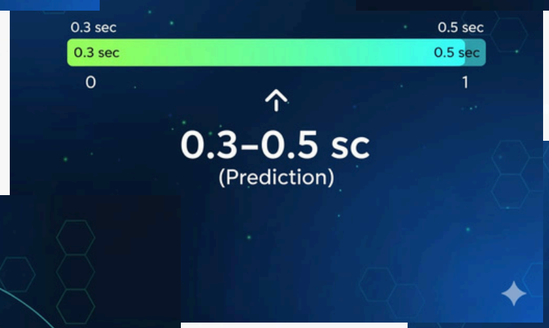


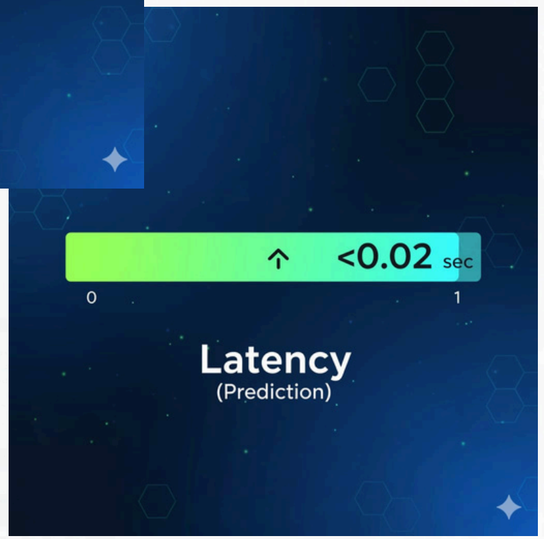


RESULTS & EVALUATION

- Testing Accuracy: 95–98%
- Real-time FPS: 30–32 FPS
- Latency: < 150 ms
- Model Size: Lightweight, fast loading
- User Feedback: Very responsive & accurate

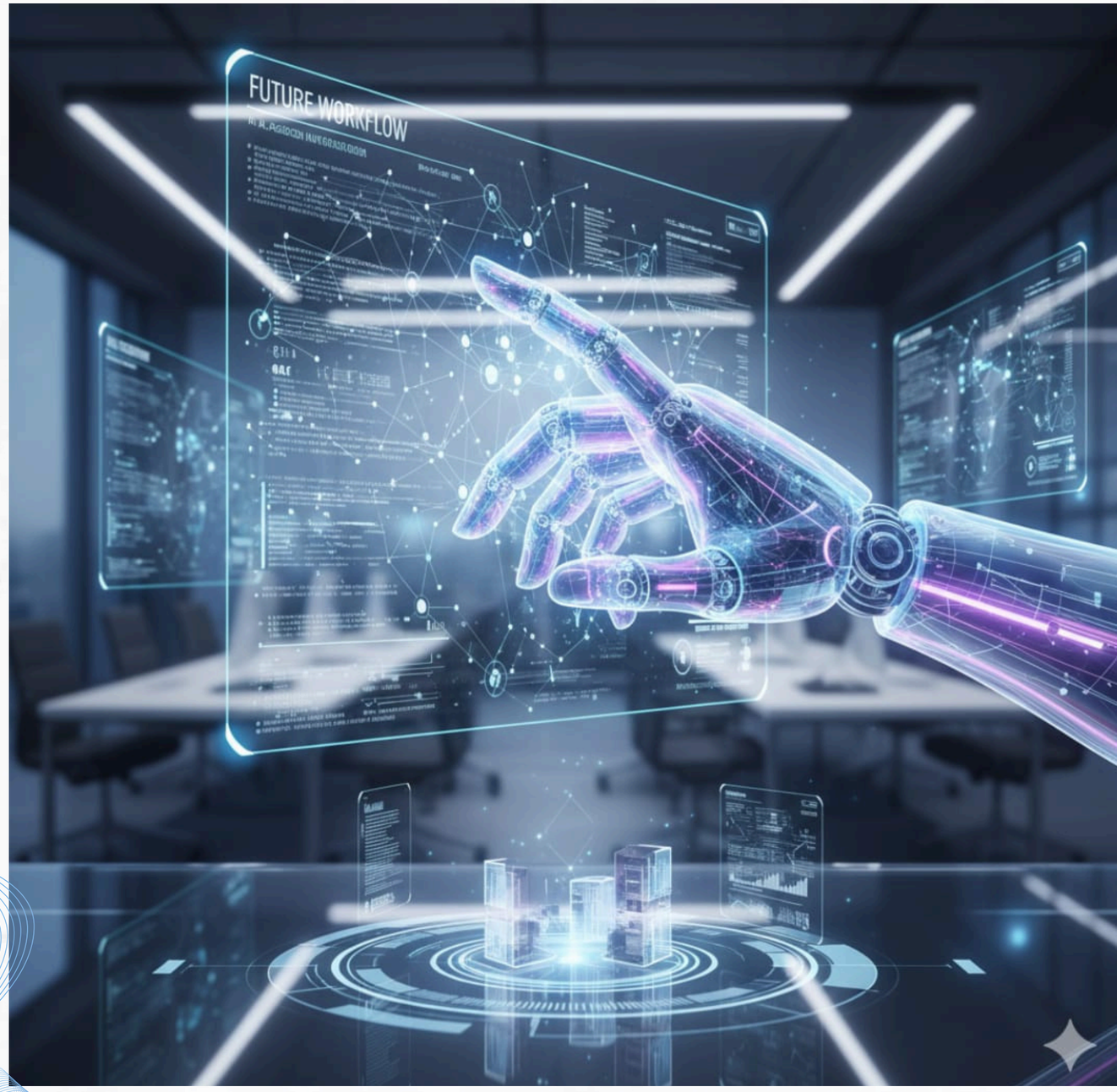






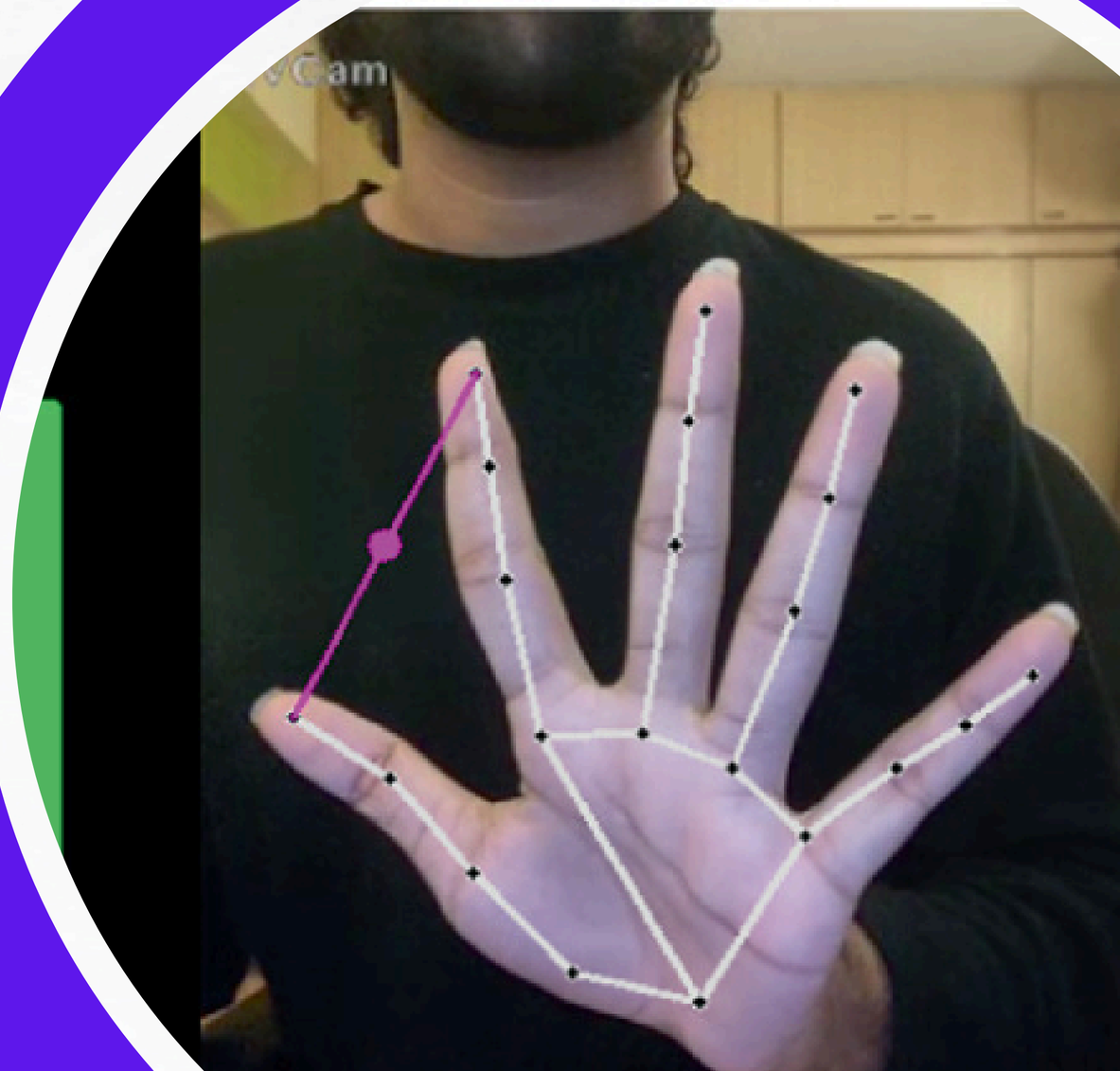
SEE OUR
FUTURE SCOPE

- Add numbers, words & dynamic gestures.
- Convert to full sign language interpreter.
- Add multi-hand / two-hand gesture support.
- Build a mobile app version.
- Integration with IoT devices.
- Add cloud-based training for scalability.



CONCLUSION

- Implemented a fully working real-time gesture recognition system.
- Uses MediaPipe + Random Forest for fast & accurate classification.
- Provides audio output for accessibility.
- Low-cost, hardware-free, and efficient interface.
- Useful for education, accessibility, and human–computer interaction.



THANK YOU