# **Table of Contents**

1. Executive Summary

2. Problem Statement

3. Project Objectives

4. Literature Review / Related Work

5. System Architecture

6. Data Description

7. Methodology

8. Implementation Details

9. Model Performance

10. Challenges and Limitations

11. Future Work

12. Conclusion

13. References

# Executive Summary

### **1.1 Project Title**

Development of an ML-Based Model for Object Detection and Holistic Body Detection

### 1.2Type of Project:

Machine Learning Model

### **1.3 Stakeholders**

Project Team Members (students)

Project Guide/Supervisor

Academic Institution (for evaluation)

End-users (for future deployment or demonstration)

### **1.4 Duration**

20th March 2025 – 29th April 2025

### **1.5 Main Objective**

The main objective of this project is to develop an efficient machine learning model for object detection and holistic body detection using MediaPipe, focusing on real-time performance with live webcam input along with static images and recorded videos. The model aims to accurately detect a variety of objects while simultaneously identifying full-body features such as pose landmarks, facial landmarks, and hand movements. By leveraging MediaPipe’s pre-trained solutions and enhancing them with custom data, the model will be capable of handling different real-world conditions like dynamic movements, varying lighting, and partial occlusions. This system is intended to provide a practical foundation for applications in areas such as human-computer interaction, fitness tracking, surveillance, and activity recognition, ensuring high accuracy and smooth, real-time execution.

# 2. Problem Statement

### **2.1 Background and Context**

In recent years, there has been significant growth in the demand for real-time computer vision systems capable of interpreting human actions and surroundings. Applications such as fitness monitoring, gesture-based control, surveillance, and virtual interaction rely heavily on robust object and body detection mechanisms. Traditional models often struggle with balancing accuracy and speed, especially when processing live inputs from webcams or high-resolution video sources. With the advancement of lightweight frameworks like MediaPipe, it has become feasible to implement high-performance models that can run efficiently even on low-resource systems.Additionally, the integration of such models into real-world systems has been limited by complex architectures and hardware requirements. Our project aims to bridge this gap by leveraging MediaPipe’s simplicity and efficiency to deliver a versatile solution that works across multiple platforms and input types.

### **2.2 Specific Problem Definition**

Despite the availability of several detection frameworks, there is a lack of integrated models that can perform both object detection and holistic human body detection (including face, pose, and hand landmarks) in real-time using diverse input sources such as webcam feeds, images, and videos. The specific challenge addressed in this project is to develop a unified ML model that performs both tasks accurately, in real-time, and under various environmental conditions such as motion, lighting variations, and occlusions. The model must be optimized for performance, responsiveness, and ease of integration into downstream applications.

### **2.3 Target Audience / Users**

The primary users of this model include developers, researchers, and students working on computer vision, human activity recognition, and interactive applications. It is also intended for use in industries focused on surveillance, virtual fitness coaching, and gesture-based control systems. Additionally, it can serve as an educational tool for institutions teaching machine learning and computer vision by demonstrating real-world implementation using lightweight frameworks like MediaPipe.

### **3. Project Objectives**

### **3.1 Primary Goals**

The primary goal of this project is to build a machine learning model capable of performing both object detection and holistic body detection (including face, hand, and pose landmarks) using MediaPipe. The model is intended to work efficiently with various input formats, such as static images, pre-recorded videos, and real-time webcam streams. It aims to classify and localize multiple objects in the input frame while simultaneously identifying key points on the human body with high accuracy and minimal latency.  
 A crucial aspect of this goal is achieving real-time performance without compromising detection quality, especially in dynamic environments. By training and fine-tuning the model using appropriate datasets, we intend to build a reliable system that can be deployed in real-world applications such as activity recognition, gesture-based control, and safety surveillance.

### **3.2 Secondary Goals**

In addition to developing the core ML model, a key secondary objective is to build a fully functional web application that integrates the trained model. This web app serves as the user interface, allowing users to interact with the model by uploading images, playing videos, or activating their webcam for real-time detection. The app is designed to process all these inputs and display the detection results directly within the browser with an interactive and intuitive user experience.  
 To ensure security and personalization, the web app includes a user login and authentication system, enabling users to securely access their session and potentially save or track their interactions with the model in future iterations. The model is deployed as a backend service within this web application, and it processes the inputs with consistent output visualization. The architecture is modular, allowing for future expansion such as user dashboards, result storage, or integration of analytics.

### **3.3 Success Criteria**

The success of the project is determined based on both model performance and system usability. For the machine learning model, accuracy will be evaluated based on detection confidence thresholds, where we are currently achieving confidence scores in the range of **0.3 to 0.4**, which is sufficient for basic detection and responsiveness while avoiding overfitting or false positives. Although the threshold is relatively low, it aligns with the goal of maintaining smooth operation in a lightweight, real-time environment.  
 To ensure smooth performance and avoid glitches or frame drops, we have configured the system to process at 2 frames per second (FPS). This allows the model to handle input in near-real time while maintaining stability and responsiveness on standard computing resources without requiring a GPU. For the web app, success is defined by the correct functioning of image, video, and webcam input modes, as well as successful user login and session handling. The application must process and display outputs with minimal latency and intuitive user interaction. Functional testing, user feedback, and measurable responsiveness will be used to evaluate overall system success.