actuators

MDPI

Article

Autonomous Robotic Platform for Precision Viticulture: Integrated Mobility, Multimodal Sensing, and AI-Based Leaf Sampling

Miriana Russo *,†, Corrado Santoro †, Federico Fausto Santoro † and Alessio Tudisco †

Department of Mathematics and Computer Science, University of Catania, 95123 Catania, Italy; corrado.santoro@unict.it (C.S.); federico.santoro@unict.it (F.F.S.); alessio.tudisco@studium.unict.it (A.T.)

* Correspondence: russo.miriana@studium.unict.it

† These authors contributed equally to this work.

Abstract

Viticulture is facing growing economic and environmental pressures that demand a transition toward intelligent and autonomous crop management systems. Phytopathologies remain one of the most critical threats, causing substantial yield losses and reducing grape quality, while regulatory restrictions on agrochemicals and sustainability goals are driving the development of precision agriculture solutions. In this context, early disease detection is crucial; however, current visual inspection methods are hindered by subjectivity, cost, and delayed symptom recognition. This study presents a fully autonomous robotic platform developed within the Agrimet project, enabling continuous, high-frequency monitoring in vineyard environments. The system integrates a tracked mobility base, multimodal sensing using RGB-D and thermal cameras, an AI-based perception framework for leaf localisation, and a compliant six-axis manipulator for biological sampling. A custom control architecture bridges standard autopilot PWM signals with industrial CANopen motor drivers, achieving seamless coordination among all subsystems. Field validation in a Sicilian vineyard demonstrated the platform's capability to navigate autonomously, acquire multimodal data, and perform precise georeferenced sampling under unstructured conditions. The results confirm the feasibility of holistic robotic systems as a key enabler for sustainable, data-driven viticulture and early disease management. The YOLOv10s detection model achieved good precision and F1-score for leaf detection, while the integrated Kalman filtering visual servoing system demonstrated low spatial tolerance under field conditions despite foliage sway and vibrations.

Keywords: computer vision; sampling system; autonomous robot; YOLO

Check for updates

Academic Editor: Zhuming Bi

Received: 19 December 2025
Revised: 12 January 2026
Accepted: 13 January 2026
Published: 2 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Modern viticulture is undergoing a critical transformation, increasingly driven by the adoption of intelligent and autonomous systems for crop management. This transition is not merely a technological evolution but a direct response to urgent economic and environmental challenges. Among these, phytopathologies represent a major threat, causing annual global losses estimated in the billions of euros [1]. These diseases compromise grapevine production both quantitatively, leading to yield reductions that may reach 30–40% in severe cases, and qualitatively [2,3] by altering the phenolic and aromatic composition of grapes and wines [4,5]. Simultaneously, growing regulatory restrictions on agrochemical use,

aligned with the European Green Deal and global sustainability goals, are accelerating the adoption of precision and low-impact agricultural practices.

Within this context, the detection of diseases plays a pivotal role. However, conventional monitoring in vineyards still relies predominantly on human visual inspection, an approach characterised by inherent subjectivity and variability among operators, as well as high labour and time costs [6]. These limitations reduce both the frequency and spatial coverage of inspections and, more critically, restrict detection to late stages of infection, when visual symptoms have already emerged and pathogens have potentially spread throughout the vineyard. Such delays make containment efforts more complex, costly, and environmentally intensive. The main goal is therefore to enable a rapid response capable of precisely targeting the affected area, preventing the spread to nearby plants. This requires innovative, high-throughput monitoring solutions far beyond the capabilities of human inspection [7].

Autonomous robotic platforms offer a transformative opportunity to overcome these constraints. Recent advances in agricultural robotics demonstrate the feasibility of ground and aerial robots for crop monitoring, canopy inspection, and targeted interventions [8–10]. These systems can enable high-frequency, high-resolution data acquisition, also known as high-throughput phenotyping, by combining continuous spatial coverage with multimodal sensing capabilities, including thermal, hyperspectral, and RGB-D modalities [11–13]. This integration enables the identification of physiological stress indicators that may appear before visible disease symptoms, providing the basis for truly proactive crop management strategies. Despite these advances, the implementation of a fully autonomous, holistic robotic system remains an open engineering challenge. Most existing research efforts in agricultural robotics focus on isolated functionalities, such as navigation [14], perception [15], or manipulation [16], without achieving seamless integration across subsystems. As a result, practical deployments often face reliability and interoperability limitations under real field conditions. Achieving a coherent architecture that synchronises mechatronic, electronic, and software components into a unified operational framework is therefore a key research frontier. A robot capable of navigating efficiently but unable to perceive or interact with its target does not constitute an effective solution.

The *Agrimet* [17] project directly addresses this gap by developing a fully autonomous robotic platform specifically tailored for vineyard environments. The system was conceived from the ground up to achieve complete operational independence from mobility to perception to manipulation and to function reliably in the challenging, unstructured conditions typical of Mediterranean vineyards. The platform is built on a tracked chassis selected for its ability to operate on slopes while minimising soil compaction. Its core is an NVIDIA Jetson Orin AGX 64GB AI–edge computing unit running a custom control software which coordinates the entire operational pipeline. The robotic platform is basically a crawler equipped with a Universal Robots UR10e [18] collaborative arm; the aim is to autonomously travel over vineyards, using a path pre-established by an operator, and perform (*i*) the sampling of visible and thermographic images of leaves, (*ii*) the gathering of leaf samples, and (*iii*) the gathering of insects present on the leaves. To achieve this goal, besides the motion system and the arm, two cameras, a RGBD Intel Realsense R435 and a FLIR A86 thermal camera, are employed, the former to perform visual imaging and to drive the arm during the leaf and insect sample gathering operations and the latter to acquire the temperature spectrum of the analysed leaves. In this robotic platform, the task of detecting and gathering leaf samples is particularly challenging since we must ensure that the end effector of the arm is properly driven towards the leaf to gather. To achieve this objective, an AI-based perception pipeline built on the state-of-the-art (SOTA) YOLOv10 deep learning model is designed, which is described in this paper. The model was trained and validated

on a proprietary dataset acquired and annotated directly in a Sicilian vineyard, ensuring robustness to variations in illumination, occlusions, and canopy geometry. The manipulator is further adapted for agricultural applications, utilising compliant control strategies to minimise tissue damage during sampling operations. A novel dual sampling workflow is also introduced, allowing for both leaf and insect collection through a custom end effector that integrates gripping and suction mechanisms. Samples are automatically stored in a carousel device that deposits specimens in ethanol-filled vials for molecular preservation. The main contributions of this work can be summarised as follows:

1. Design and implementation of a multimodal perception system enabling high-throughput phenotyping in vineyard environments.
2. Application of AI-based visual servoing for real-time leaf localisation during robotic manipulation.
3. Adaptation of an industrial collaborative robot to compliant manipulation tasks in agricultural settings.
4. Design and validation of a dual biological sampling workflow for georeferenced field data acquisition.

The proposed platform demonstrates the potential of integrated perception, control, and manipulation strategies toward autonomous robotic solutions for data-driven viticulture.

The remainder of this paper is organised as follows: Section 2 reviews related work in agricultural robotics and vision-based crop monitoring; Section 3 proposes an analysis of the work environment and the requirements; Section 4 details the proposed system architecture and describes in details the mechanical, electronic, and software subsystems; Section 5 describes the pipeline that the system follows and the interaction between different modules and implementation; Section 6 presents the experimental validation in a vineyard environment; and Section 7 summarises the main findings and outlines future research directions.

2. Related Works

The deployment of autonomous robotic platforms in agricultural environments has gained significant attention as a solution to address labour shortages, improve monitoring efficiency, and enable precision agriculture. Ground-based robots equipped with advanced sensors provide continuous, high-throughput data collection capabilities that surpass traditional manual inspection methods. Recent developments demonstrate tracked robotic systems operating in vineyards for various tasks, including surveillance, spraying, and crop monitoring. The VINBOT project [8] pioneered autonomous vineyard monitoring by developing an all-terrain robot capable of capturing and analysing vineyard images through cloud computing to determine yield and canopy features. Similarly, the Bacchus Long-Term [9] deployment demonstrated the feasibility of persistent robotic operation across seasonal changes in vineyard environments. However, most existing systems focus on isolated functionalities such as navigation or perception without achieving seamless integration across all subsystems.

The challenge of developing fully autonomous, holistic robotic platforms that coordinate mobility, perception, and manipulation remains an active research frontier, particularly for complex tasks like biological sampling in unstructured vineyard environments.

Machine learning and deep learning approaches have revolutionised crop disease detection [19,20] by enabling automated, accurate identification of plant diseases from visual imagery. Convolutional Neural Networks (CNNs) have emerged as the dominant approach, with architectures such as VGG [21,22], ResNet [23,24], and DenseNet [25] achieving detection accuracies exceeding 95% on labelled crop disease datasets. The progression from traditional handcrafted feature extraction to automated deep learning-

based feature learning has significantly improved the robustness and scalability of disease detection systems [26]. Recent advances incorporate real-time object detection models like Faster R-CNN [27,28], YOLO [29], and SSD [30,31], which not only identify but also localise disease symptoms within crop canopies, enabling targeted intervention strategies. Vision Transformers (ViTs) [32] represent an emerging approach that leverages self-attention mechanisms to capture long-range dependencies in high-resolution plant images [33,34], showing promising improvements over CNNs for complex disease classification tasks. The integration of UAV-based aerial imaging [35,36] with deep learning has expanded disease monitoring capabilities across large agricultural landscapes.

Despite these advances, challenges remain in achieving robust performance under varying illumination conditions, handling occlusions in dense canopies, and generalising trained models across different crop varieties and environmental conditions.

Multimodal sensing combining RGB, depth, and thermal data has emerged as a powerful approach for comprehensive plant health assessment and phenotyping [11]. RGB-D sensors, particularly the Intel RealSense series [37], provide synchronised colour and depth information enabling 3D reconstruction of plant structures and accurate measurement of morphological traits. Thermal imaging complements visual data by detecting physiological stress indicators such as elevated leaf temperature associated with water stress or disease infection, often before visible symptoms appear [12]. The fusion of RGB-D and thermal modalities addresses limitations inherent to single-sensor approaches: RGB-D sensors excel at structural characterisation but may struggle with subtle physiological changes. In contrast, thermal cameras detect stress but lack the spatial resolution for precise localisation.

Recent studies demonstrate that combining these modalities through machine learning frameworks significantly improves prediction accuracy for soil moisture estimation, plant water stress detection, and disease monitoring. However, sensor fusion introduces technical challenges, including calibration and registration between sensors with different fields of view and resolutions, data synchronisation, and computational demands for real-time processing. Advanced approaches employ calibration techniques to align multimodal data streams and machine learning models to extract complementary features from each modality for robust agricultural monitoring applications [13].

The YOLO family [38] of object detection algorithms has become increasingly popular for agricultural applications due to its balance between detection accuracy and real-time inference speed. YOLOv10 [39], the latest iteration, introduces significant architectural innovations, including NMS-free training through consistent dual assignments, eliminating post-processing bottlenecks that limited previous versions. The architecture employs an efficiency-driven design, featuring lightweight classification heads, spatial-channel decoupled downsampling, and a rank-guided block design, to reduce computational overhead while maintaining detection performance. For accuracy enhancement, YOLOv10 incorporates large-kernel convolutions to expand receptive fields and partial self-attention modules for global representation learning with minimal computational cost.

Agricultural robotics has successfully deployed YOLO variants for real-time plant disease detection [40], with YOLOv5 and YOLOv8 demonstrating detection accuracies exceeding 90% for various crop diseases. The nano and small variants of YOLOv10 are particularly suited for edge computing deployment on resource-constrained agricultural robots [41], achieving superior performance–efficiency trade-offs compared to previous versions. Studies demonstrate YOLOv10's effectiveness for small object detection in challenging conditions [42], making it ideal for identifying individual leaves and disease symptoms in dense canopy environments [43]. The model's end-to-end architecture enables deployment on embedded systems with reduced latency, supporting real-time visual servoing applications for robotic manipulation in agriculture. Beyond discriminative

deep learning models for perception, recent research has explored the role of generative artificial intelligence in digital agri-food systems. Generative AI techniques have been investigated for tasks such as data augmentation, synthetic dataset generation, decision support, and knowledge extraction in precision agriculture workflows [44]. While these approaches primarily operate at a higher semantic and data management level, the present work focuses on real time perception and control for autonomous robotic manipulation in vineyard environments.

Collaborative robots (or cobots) are transforming agricultural automation by enabling safe human–robot interaction and performing delicate manipulation tasks that require precision and adaptability. The deployment of industrial collaborative manipulators such as Universal Robots' UR series [45] in agricultural settings demonstrates their capability to execute complex tasks, including selective harvesting, precision spraying, and biological sampling. These robots feature integrated force–torque sensing that enables compliant interaction with plants, minimising mechanical damage during sampling operations while maintaining high repeatability.

Agricultural manipulation presents unique challenges compared to industrial applications, including unstructured environments, high variability in target geometry and position, and the need to handle delicate biological materials. Research has explored human–robot collaboration strategies [46] that combine robotic efficiency with human cognitive skills, particularly for tasks requiring adaptability and decision-making in variable field conditions [16].

The integration of cobots with vision-guided control systems [47] enables precise localisation and approach trajectories for leaf sampling and targeted interventions. Studies on vineyard operations [10] demonstrate that collaborative mobile manipulators can assist workers by reducing biomechanical stress and improving task efficiency in challenging terrains. However, significant research gaps remain in developing robust manipulation strategies for variable lighting conditions, handling occlusions in dense canopies, and ensuring sample quality consistency across diverse agricultural scenarios.

Visual servoing integrates real-time visual feedback to guide robotic manipulator motion, enabling precise positioning in dynamic and unstructured environments. The Kalman filter has emerged as an effective approach for visual servoing control [48], providing optimal state estimation by recursively combining noisy sensor measurements with predictive system models. In visual servoing applications, Kalman filters estimate joint-space positioning errors and filter noise from image-based measurements, resulting in smoother trajectories and improved convergence compared to traditional Gauss–Newton approaches. Studies demonstrate that Kalman filter-based visual servoing achieves superior performance in high-noise scenarios [49], which are common in agricultural environments with varying illumination and dynamic backgrounds. The filter's recursive nature and ability to handle uncertainty make it particularly suitable for tracking moving or swaying targets [50] such as leaves affected by wind. Dynamic visual servoing systems [51] incorporating Kalman filtering can estimate both position and velocity states from image data alone, eliminating the need for additional velocity sensors. For uncalibrated visual servoing, where precise camera and robot calibration parameters are unavailable, Kalman filtering combined with online Jacobian estimation [52] enables effective control without extensive system modelling. Agricultural applications benefit from these robust control approaches, as they enable consistent performance despite environmental variability, sensor noise, and approximate system models that characterise real-world field deployments.

Autonomous navigation based on GPS waypoint following has become a fundamental capability for agricultural ground robots operating in open-field environments. GPS-based navigation systems enable robots to autonomously traverse predefined paths or

reach specific locations without continuous human intervention, essential for repetitive agricultural tasks such as monitoring, spraying, and data collection.

Modern implementations [53] integrate high-precision GPS receivers with inertial measurement units (IMUs) and magnetometers to achieve decimetre-level positioning accuracy and reliable heading determination. The RoboNav [54] system demonstrates that cost-effective dual GPS configurations combined with Gaussian Sum Filters can achieve average positioning errors of 0.2 m and heading errors of 0.2 degrees in vineyard environments. Autonomous ground vehicle navigation architectures typically employ hierarchical control structures with high-level mission planning, mid-level path planning, and low-level trajectory tracking controllers.

Challenges in agricultural navigation include GPS signal degradation under dense canopy cover, the need for obstacle detection and avoidance in unstructured environments, and maintaining stable control on sloped and irregular terrain [14]. Advanced systems integrate LiDAR or camera-based perception for local obstacle detection, enabling reactive navigation around unexpected obstructions while maintaining global GPS-guided path following [55]. Recent developments in SLAM technology provide robust localisation even in GPS-denied areas, though computational requirements remain a consideration for resource-constrained agricultural platforms.

Precision agriculture leverages advanced technologies, including sensors, artificial intelligence, and autonomous systems, to enable site-specific crop management, optimising resource use while minimising environmental impact. Automated disease monitoring systems address limitations of traditional visual inspection methods, labour intensity, subjectivity, and late detection by providing continuous, objective assessment of crop health. The integration of IoT, machine learning, and remote sensing creates intelligent operational systems capable of real-time field management and predictive analytics for disease forecasting.

Multimodal sensing approaches [12] combining visible, multispectral, thermal, and depth imaging enable the early detection of pre-symptomatic stress indicators such as altered leaf temperature, spectral reflectance, or structural changes. AI-driven disease detection systems [6] achieve high accuracy (>90%) for multiple crop diseases, supporting automated monitoring and decision systems that optimise pesticide application timing and placement.

Despite technological progress, challenges remain in translating research advances into widespread agricultural practice, including high implementation costs, data privacy concerns, limited rural digital infrastructure, and the need for farmer training. Future developments in edge AI, biosensors, and interoperable data platforms promise to make precision disease management more accessible and effective, supporting sustainable agriculture in the face of climate change and increasing production demands.

The review of the existing literature reveals that while significant advances have been made in individual components of agricultural robotics, including autonomous navigation, computer vision, deep learning for disease detection, and robotic manipulation, a critical research gap persists regarding the holistic integration of these technologies into fully autonomous, field-deployable systems for vineyard monitoring. Most existing platforms focus on isolated functionalities such as navigation or perception without achieving seamless coordination across mobility, sensing, and manipulation subsystems. This fragmentation limits practical deployment under real-world vineyard conditions, where complex terrain, variable environmental factors, and unstructured canopy architectures demand robust, integrated solutions.

The present work addresses these gaps by developing a fully integrated autonomous robotic platform that unifies tracked mobility for slope operation, multimodal perception

using RGB-D and thermal sensors, AI-based visual servoing with YOLOv10 for real-time leaf localisation, collaborative robotic manipulation for compliant biological sampling, and a complete georeferenced data acquisition workflow. By demonstrating the feasibility of holistic system integration tailored specifically for Mediterranean vineyard environments, this research advances the state-of-the-art in precision viticulture and establishes a foundation for next-generation agricultural robots capable of enabling data-driven, sustainable crop management strategies in response to evolving climate and regulatory pressures.

3. Operating Environment and Goals

The Agrimet initiative recognises that modern viticulture faces unprecedented challenges: climate change is altering disease patterns and introducing new pathogens, labour costs and availability constrain frequent monitoring, and increasing consumer and regulatory demands for sustainable practices require minimising pesticide applications. These factors necessitate a paradigm shift toward technology-driven, data-intensive vineyard management.

3.1. Vineyard Environment Characteristics

Understanding the operational environment is crucial for designing an effective robotic system. Mediterranean vineyards, particularly those in Sicily, present unique challenges that directly influence system requirements.

As illustrated in Figure 1, vineyards exhibit structured but variable characteristics:

- **Row architecture**: Inter-row spacing typically ranges from 2.0 to 3.5 m, with plant spacing within rows of 0.8 to 1.5 m. These dimensions constrain robot size and manoeuvrability requirements. Accordingly, the system employs a compact tracked chassis designed to fit comfortably within inter-row spacing while maintaining stability. The UR10e collaborative arm's 1.3 m reach ensures access to leaves distributed across the 0.8–1.5 m intra-row spacing, while its 0.05 mm repeatability guarantees consistent sampling protocols despite varying plant geometries.
- **Terrain morphology**: Sicilian vineyards often occupy hillside locations with slopes reaching 15–25%, featuring irregular surfaces, exposed rocks, and variable soil compaction. These conditions demand robust locomotion systems with high ground clearance and traction. As detailed in the subsequent analysis, the tracked platform with dual 1.5 kW brushless motors and field-oriented control enables stable navigation on inclined terrain while minimising soil compaction. GPS/IMU fusion (Pixhawk 4 autopilot with magnetometer integration) provides decimetre-level positioning accuracy necessary for systematic coverage across variable topography.
- **Canopy structure**: Training systems (primarily Guyot and cordon) create complex three-dimensional canopy architectures where target leaves may be occluded, requiring sophisticated vision algorithms and flexible robotic manipulation. The system directly addresses this complexity through multimodal perception: the RGB-D camera (Intel RealSense D435, 0.3–3.0 m working range at 30 Hz) provides synchronised colour and depth data enabling 3D leaf localisation in cluttered environments, while Kalman filtering with constant velocity motion prediction handles temporary occlusions (up to 4 s) by estimating leaf position during wind-induced sway. This approach achieves 3 cm spatial tolerance despite foliage obstruction and dynamic motion. The compliant six-axis manipulator with integrated force–torque sensing enables selective leaf sampling while minimising tissue damage during insertion into occluded canopy regions.
- **Ground conditions**: Soil moisture varies significantly with season and irrigation, ranging from hard, compacted surfaces to soft, muddy conditions that challenge traction and stability. The tracked architecture addresses this variability by distributing

robot weight across a large surface area, reducing ground pressure compared to wheeled platforms.

- **Obstacles**: Environmental factors: Operations must tolerate ambient temperatures of 5–40 °C, high humidity levels, direct solar radiation, wind, and occasional precipitation. Wind-induced leaf sway is directly addressed through Kalman filtering with state prediction (incorporating accelerometer data), enabling robust target tracking despite dynamic foliage motion. Variable illumination, from direct sunlight to backlit conditions in dense canopy, necessitates multimodal sensing: the thermal camera (Teledyne FLIR A68, 50 mK sensitivity) detects physiological stress indicators (elevated leaf temperature) independent of visible-light conditions, enabling early disease detection that RGB imaging alone cannot achieve. The requirement to generalise perception across varying environmental conditions informed the decision to train the YOLOv10 model on a proprietary dataset acquired directly in a Sicilian vineyard, capturing realistic variations in illumination, occlusion patterns, and canopy geometry. Obstacles for vineyard infrastructure include support posts typically every 5–8 m, irrigation lines, ground anchors, and occasionally fallen branches or maintenance equipment. The Pixhawk 4 autopilot integrates LiDAR-assisted obstacle detection with GPS waypoint following, enabling the system to autonomously navigate around infrastructure while maintaining mission-level path adherence during unsupervised operations.
- **Environmental factors**: Operations must tolerate ambient temperatures of 5–40 °C, high humidity levels, direct solar radiation, wind, and occasional precipitation.



Figure 1. Typical vineyard layout showing row spacing and plant arrangement.

These environmental constraints establish fundamental design parameters for all robotic subsystems, from mechanical chassis design to sensor selection and control algorithms. The integrated system architecture demonstrates how individual environmental challenges map to specific technical solutions: tracked mobility for slope operation and soil protection; GPS/IMU fusion for decimetre-level positioning despite terrain variability; RGB-D + Kalman filtering for robust perception despite canopy occlusions and wind-induced motion; thermal imaging for stress detection independent of illumination conditions; and local dataset training for environmental generalisation. This principled, requirement-driven design approach ensures that every subsystem addresses concrete operational constraints, maximising field reliability in unstructured vineyard environments.

The lower centre of gravity and greater adhesion, characteristic of a tracked system, constitute key elements for operating on slopes or uneven surfaces. In vineyards, where rows are often arranged along slopes, the ability to face inclinations without risk of tipping is crucial. Conversely, using wheels could more easily lose grip and cause lateral oscillations that put both the robot and plants at risk. Tracks ensure continuous contact with the ground, improving the ability to overcome small obstacles such as stones or protruding roots and reducing the risk of slipping even in difficult conditions such as mud, wet grass, or sandy terrain. The robot can therefore operate reliably throughout the year, without being limited by weather conditions.

3.2. Developing an Autonomous Robotic System

The goal is to design, implement, and validate an autonomous robotic platform specifically engineered for systematic biological sampling in vineyard environments. This system aims to revolutionise disease surveillance and agronomic monitoring by providing the following:

- **Enhanced efficiency**: Automated sampling reduces labour requirements by up to 70–80% compared to manual methods, enabling more frequent and comprehensive monitoring.
- **Improved precision**: Computer vision-based targeting and robotic manipulation ensure consistent, reproducible sampling protocols, eliminating human variability.
- **Comprehensive spatial coverage**: GPS-guided navigation enables systematic sampling across entire vineyard blocks, creating high-resolution disease distribution maps.
- **Continuous operation capability**: The system can operate during extended periods, including conditions (early morning, late evening) when human inspection is impractical.
- **Data integration**: Georeferenced samples and sensor data feed directly into geographic information systems (GISs) and decision support tools.

Beyond immediate disease monitoring applications, the developed platform serves as a research tool for studying pathogen epidemiology, vector population dynamics, and the efficacy of various disease management strategies under real-world conditions.

3.3. Design Assumptions and Operational Constraints

The development and deployment of the proposed system rely on several key assumptions that define its operational scope and limitations:

- Vineyard structure: The system requires inter-row spacing at least 2.0 m to ensure safe navigation and manoeuvrability of the tracked platform. Vineyards with narrower row configurations are not compatible with the current mechanical design.
- Weather conditions: Operations are conducted during dry weather conditions. Rain or heavy fog may compromise traction, sensor performance, and sample quality, limiting system deployment.
- GPS availability and accuracy: The system requires continuous GPS coverage with minimal positioning error for autonomous navigation. Standard GPS receivers are employed, though RTK GPS can optionally be integrated for enhanced accuracy in demanding applications. Areas with significant GPS signal obstruction.
- Lighting conditions: The vision system operates effectively under diffuse daylight conditions. While data augmentation techniques improve robustness to illumination variations, insufficient lighting (e.g., heavy cloud cover, twilight, darkness) significantly degrades leaf detection accuracy, as the neural network requires adequate visual information to identify target features.
- Operational hours: Given the lighting requirements, the system is validated for daylight operations (typically 7:00 a.m.–19:00 p.m. depending on season), when illumi-

nation is sufficient for reliable computer vision performance and human supervision is practical.

4. System Architecture

The platform integrates multiple interconnected subsystems into a coherent autonomous robotic architecture capable of independent operation in unstructured vineyard environments. The overall system design follows a hierarchical modular approach with three principal layers: (see Section 4.2) a locomotion and navigation layer managing autonomous mobility, (see Section 4.3) a biological sampling layer coordinating perception and manipulation, and (see Section 4.4) a software layer orchestrating subsystem coordination and mission execution. This section details the architectural integration strategy and principal component interactions.

4.1. Subsystem Organisation

The proposed architecture emphasises functional modularity and clear separation of concerns. At the foundation resides a tracked mobile platform equipped with brushless motors and industrial-grade motor drivers, providing autonomous mobility across challenging vineyard terrain. A Pixhawk 4 autopilot manages autonomous navigation through GPS-based waypoint following and IMU fusion. To bridge the communication protocol mismatch between the autopilot's PWM outputs and motor drivers' CANopen interface, a custom middleware layer on an STM32L4KC microcontroller performs real-time PWM-to-CAN translation.

The biological sampling subsystem comprises a Universal Robots UR10e collaborative arm serving as the primary manipulation platform, equipped with customised end-effectors enabling both leaf sampling and arthropod collection. Two multimodal vision sensors—an Intel RealSense D435 RGB-D camera and a Teledyne FLIR A68 radiometric thermal camera—enable perception-guided manipulation through a YOLOv10 deep learning model trained on vineyard-specific grapevine leaf data.

Orchestrating these heterogeneous components is a unified software control framework, executing on an NVIDIA Jetson Orin AGX edge AI processor. This framework abstracts hardware-specific implementation details through modular software components, enabling the coordinated execution of autonomous missions spanning navigation, target detection, manipulation, and sample preservation. For the realisation of the robotic unit, the architecture is divided into three primary subsystems, as represented in Figure 2, each responsible for distinct functional domains.

4.2. Locomotion and Autonomous Navigation System

This subsystem provides mobility and autonomous navigation capabilities, comprising the following:

- **Dual brushless electric motors**: High-torque, low-speed motors (rated at 36 V, 1.5 kW each) provide independent drive to left and right tracks, enabling skid-steering manoeuvres with zero-radius turning capability. Brushless architecture ensures high efficiency (>85%), minimal maintenance requirements, and excellent thermal characteristics for extended operations.
- **Motor driver controllers (Zapi BLE0)**: Industrial-grade motor drivers implement field-oriented control (FOC) for precise torque regulation, provide comprehensive safety features (overcurrent protection, thermal management, ground fault detection), and expose CANopen communication interfaces for high-level control integration.
- **Flight controller autopilot (Pixhawk 4)**: Originally designed for aerial vehicles, this autopilot has been adapted for ground vehicle applications, providing GPS/GNSS

positioning, inertial measurement unit (IMU) data fusion, magnetometer-based heading determination, and mission management capabilities. The platform supports MAVLink protocol for standardised communication with ground control stations.

- **PWM-to-CAN (STM32L4KC)**: This microcontroller serves as the critical interface between the PWM-based autopilot outputs and the CANopen-based motor drivers, translating high-level navigation commands into differential drive motor commands while implementing real-time control loops for trajectory tracking.

The integration of these components enables autonomous waypoint navigation, dynamic obstacle avoidance, and return-to-home functionality essential for unsupervised vineyard operations.

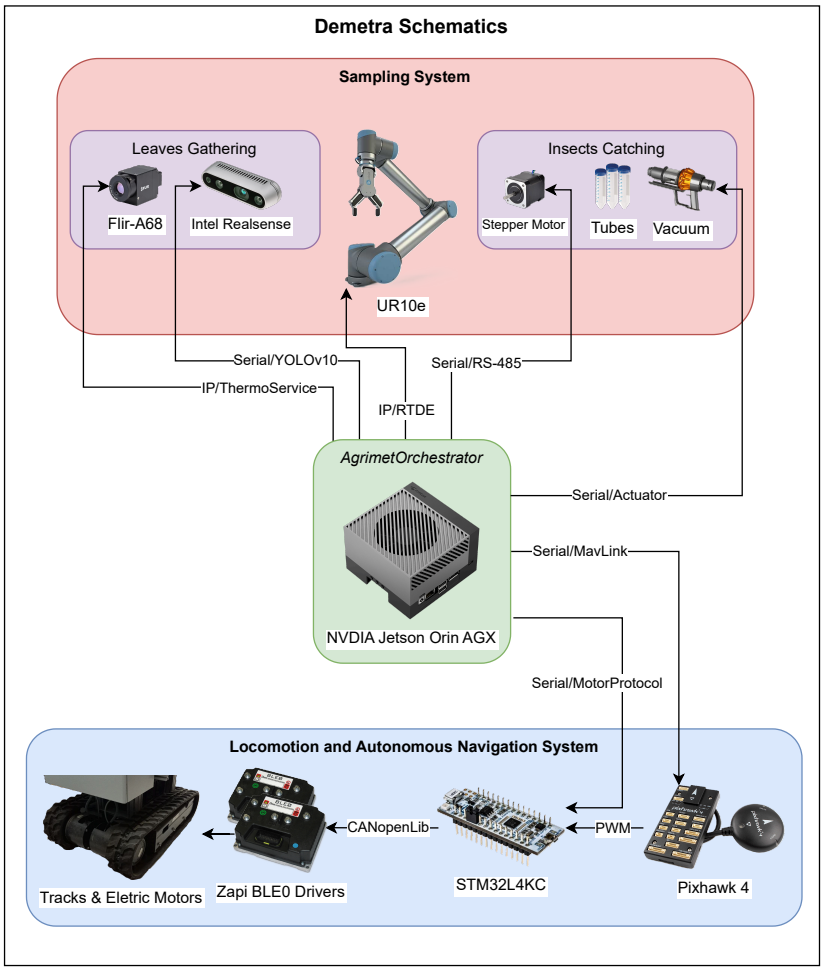


Figure 2. Robot architecture diagram showing major subsystems and their interconnections.

4.3. Sampling System

The sampling subsystem, shown in Figure 3, handles all perception and manipulation tasks required for biological sample collection:

- **RGB-D vision sensor (Intel RealSense D435)**: Provides synchronised colour and depth imaging at 640 × 480 resolution and 30 Hz frame rate, enabling 3D perception of the vine canopy. The active stereo depth technology operates reliably across varying lighting conditions and offers a working range of 0.3–3.0 m, suitable for close-range plant inspection.
- **Thermal imaging camera (Teledyne FLIR A68)**: Captures radiometric thermal data at 640 × 480 resolution with <50 mK thermal sensitivity, enabling detection of plant stress, disease symptoms, and environmental anomalies invisible to RGB cameras.

Thermal imaging can reveal pre-symptomatic disease indicators through localised temperature variations.

- **Collaborative robotic arm (Universal Robots UR10e)**: Six-axis manipulator with 1.3 m reach, 12.5 kg payload capacity, and ±0.05 mm repeatability provides the dexterity required for selective leaf sampling. The cobot architecture includes integrated force–torque sensing for compliant interaction with plants, minimising mechanical damage during sampling operations.
- **End-effector tools**: Custom-designed gripper for leaf collection and vacuum aspiration nozzle for insect capture, mountable on the robotic arm's tool changer interface. The gripper employs gentle pneumatic actuation to avoid crushing delicate leaf tissues.
- **Sample management system**: Automated carousel mechanism containing 20+ individual sample vials prefilled with preservation solution (70% ethanol for entomological specimens). A stepper motor-driven indexing system positions each vial sequentially under the sample deposit point, ensuring proper sample separation and traceability.

The sampling system's modular design enables the independent operation of its vision, manipulation, and storage subsystems, facilitating troubleshooting and future upgrades.

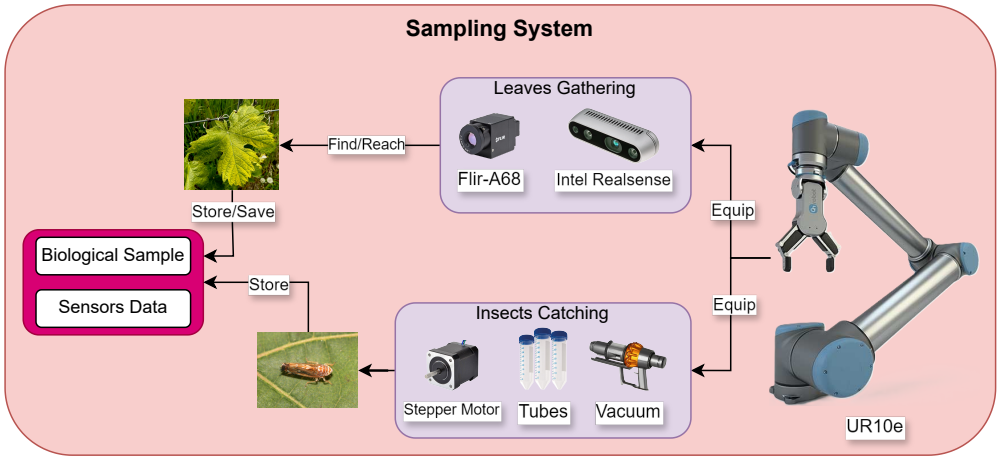


Figure 3. Sampling system operative details.

4.4. Control Software Architecture

The software architecture, shown in Figure 4, emphasises modularity, maintainability, and extensibility, allowing for future integration of additional sensors, control algorithms, or analysis tools without requiring fundamental system redesign:

- **AgrimetOrchestrator**: The main coordinator, composed of various submodules, exposes a comprehensive set of HTTP API endpoints that enable remote user interaction through custom-developed client applications. These APIs facilitate real-time system monitoring, remote control, and dynamic configuration adjustment, thereby supporting integration with graphical user interfaces (GUIs), operational dashboards, and supervisory control systems.
- **Motors Module**: Establishes and maintains a serial connection with the STM32L4KC microcontroller and runs a persistent background thread to provide three core capabilities: real-time motor telemetry acquisition (including operational status, current consumption, and performance metrics); motor control operations such as activation, deactivation, and automated calibration of tracked drive motors; and data logging management for enabling or disabling comprehensive performance logs for diagnostics and analysis.

- **RealSense Module**: Provides key functionalities such as video pipeline management for initialising, starting and stopping real-time capture streams, synchronised acquisition of RGB and depth frames at configurable resolutions and frame rates, and coordinate transformation between 2D image-space pixels and 3D real-world metric coordinates using depth sensor calibration parameters.
- **FlirA68 Module**: Provides snapshot functionality for the thermal camera.
- **VisionController**: Performs real-time grapevine leaf detection using an optimised YOLOv10 model. It processes RGB frames from the RealSense camera, applies target selection logic, and uses Kalman filtering to stabilise 3D leaf coordinates, ensuring robust tracking under dynamic vineyard conditions.
- **ArmController**: Manages high-level manipulation tasks for the UR10e robotic arm via the RTDE protocol. It integrates vision-based spatial data to execute autonomous leaf and insect sampling workflows.
- **StepperController**: Handles carousel homing, vial indexing, and position reporting, ensuring reliable sample handling.
- **Pixhawk Module**: Manages autonomous navigation using MAVsdk and MAVLink. It establishes autopilot communication, uploads waypoint missions, monitors telemetry, and coordinates sampling operations with other modules during flight.

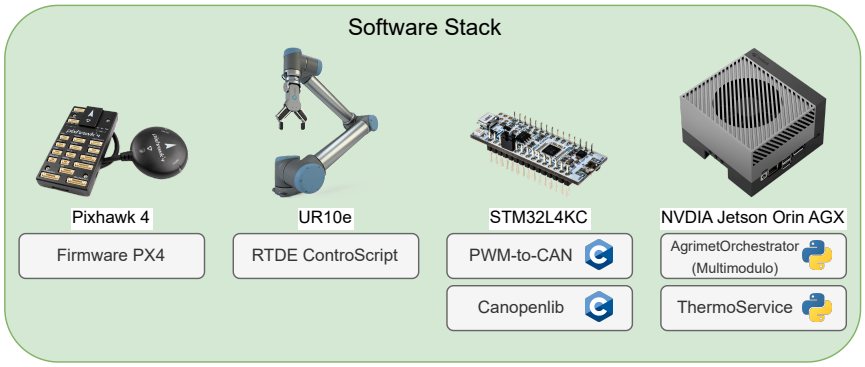


Figure 4. Software distribution over the components.

4.5. System Interconnection

The interconnection system is shown in Figure 5. NVIDIA Jetson Orin AGX serves as the computational core, providing the processing capability necessary to execute the complete operational pipeline. This platform communicates with various hardware components via multiple communication interfaces and protocols, enabling seamless integration across heterogeneous subsystems.

4.5.1. Locomotion and Autonomous Navigation Subsystem Connectivity

- The STM32L4KC microcontroller interconnects the ZAPI BLE0 motor drivers and the Pixhawk 4 autopilot through the PWM-to-CAN translation process.
- Both the STM32L4KC microcontroller and the Pixhawk 4 autopilot communicate with the central Jetson Orin AGX processor via serial interface protocols.

4.5.2. Biological Sampling Subsystem Connectivity

- The Universal Robots UR10e collaborative arm and the Teledyne FLIR A68 thermal camera communicate with the central processor through local area networking, established via integrated router and network switch hardware.
- The Intel RealSense D435 RGB-D camera and the stepper motor control board communicate with the central processor via serial interface protocols.

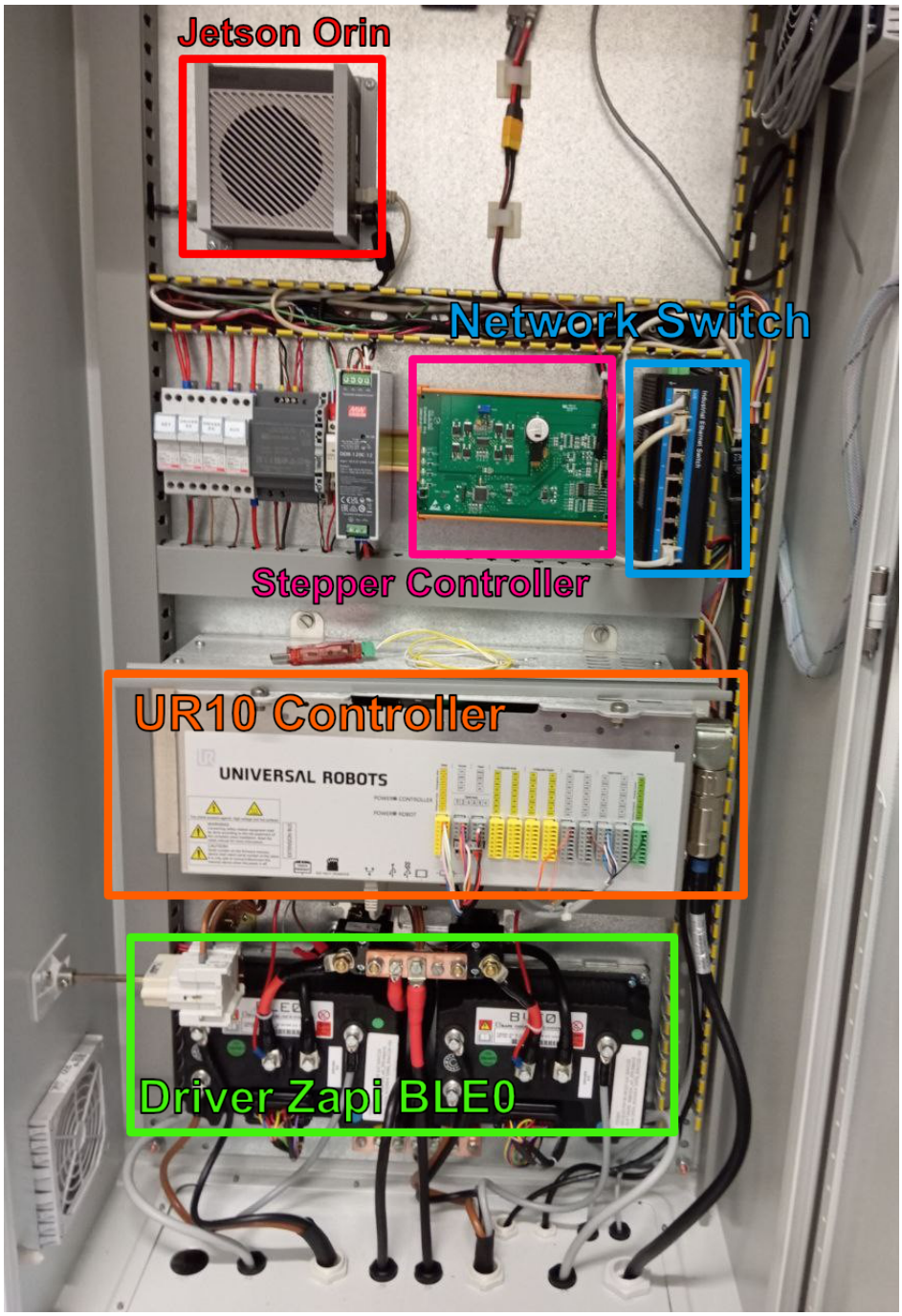


Figure 5. Interconnected components.

5. Leaf Detection and Sampling Pipeline

Harvesting leaves consists of implementing a visual servoing system that receives RGB-D frames as input, performs target leaf detection and tracking, and provides as output the 3D coordinates of the selected leaf for robotic arm control.

5.1. Operational Workflow

The complete sampling procedure, visually summarised in Figure 6, consists of the following sequential steps:

- The robotic arm starts from a view position (fixed at design time);
- The stereo camera acquires the image and searches for a good choice for a leaf;
- Once a leaf is targeted, the arm approaches and follows it with the camera;

- When the tool of the robotic arm is at the correct distance, a thermal photo is taken with the thermal camera;
- Next, the robotic arm goes forward to take a sample of the leaf;
- After that, the robotic arm positions the vacuum nozzle near a target leaf or vine structure;
- Vacuum aspiration is activated to capture insects;
- Specimens are transported through the duct and deposited into the current vial;
- The stepper motor rotates the carousel to position the next vial.

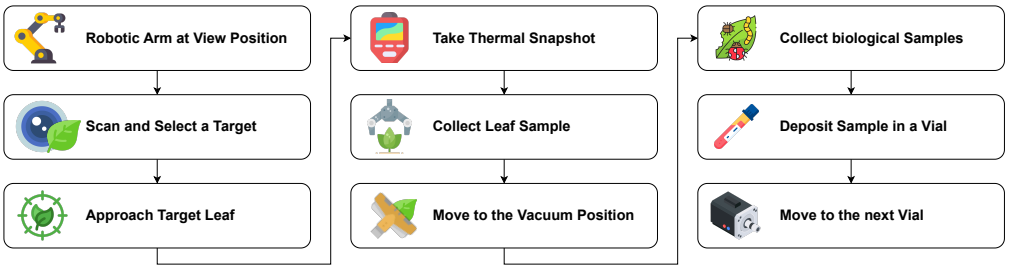


Figure 6. Operational workflow of the sampling system.

Once a target leaf has been identified, the system must ensure continuity and robustness of perception during the robotic arm motion. In outdoor agricultural environments, leaf detection can be affected by partial occlusions, sudden illumination changes, motion blur, or rapid movements of vegetation caused by wind. To address these challenges, the pipeline incorporates a target selection and tracking strategy that allows the system to maintain focus on the selected leaf throughout the approach phase.

When multiple candidate leaves are detected in the scene, a selection logic is applied to identify the most suitable target for sampling. This selection is based on detection confidence and spatial consistency, favouring leaves that present higher reliability and geometric suitability for interaction. Once a target is selected, the system enters a tracking mode in which subsequent detections are evaluated with respect to the previously estimated leaf position, ensuring temporal coherence and preventing undesired switches between nearby objects.

In cases where the target leaf is temporarily not detected due to occlusions or sensor noise, the pipeline does not immediately discard the target. Instead, a Kalman filter is employed. This recursive algorithm estimates the internal state of a linear dynamic system from a series of noisy measurements, allowing the prediction of the most probable leaf position based on its motion history. This approach allows the robotic arm to continue its motion smoothly, avoiding abrupt stops or erratic corrections that could compromise the stability of the sampling operation.

To further improve positional stability, temporal filtering techniques are applied to the estimated three-dimensional coordinates of the target leaf. These mechanisms reduce measurement noise and smooth the perceived leaf trajectory over time, contributing to a more controlled and reliable visual servoing behaviour. As a result, the robotic arm can approach the target leaf with progressive and stable movements, maintaining the desired operating distance and orientation required for thermal acquisition and sampling tasks.

Overall, the proposed pipeline enables robust perception-to-action coupling by integrating detection, depth-based localisation, tracking, and stabilisation within a unified control framework. This design ensures reliable leaf sampling even under challenging environmental conditions, while preserving real-time performance and system responsiveness.

5.2. Building a Leaf Detector

5.2.1. Dataset Definition

To achieve a better success rate for recognising wine leaves, we decided to fine-tune a YOLO model using a custom dataset based on real footage of Sicilian variant wine plants. Acquisitions were conducted in a vineyard located in the territory of Bronte, in the province of Catania (Italy), to include realistic environmental situations, such as natural lighting and the type of vine typical of the operating area (an example of this environment is shown in Figure 7).

The acquisition was performed with an Intel Realsense D435 camera, the same model as the one equipped on the robotic arm, to ensure coherence of focal parameters between training data and real operational acquisitions. Videos were acquired with 640 × 480 pixel resolution at 30 FPS, keeping the camera in motion to simulate the robot's dynamic approach to targets, obtaining observations with variations in scale, rotation, and perspective that characterise the visual servoing task, the robotic arm's approach to the leaf.



Figure 7. Example of a frame we can find inside the dataset.

By extracting about 500 distinct frames from the videos, a dataset characterised by good diversity was obtained, documenting leaves at various development stages, with various lighting conditions (direct, diffuse, backlight), in the presence of partial occlusions due to other leaves or vine structure elements.

5.2.2. Labelling

The annotation of training set images constitutes a critical phase requiring precision and consistency to ensure optimal model performance. For this reason, the robotic system was created to select and track a maximum of four leaves per operating cycle, making the annotation of partially visible elements possible. A selective annotation strategy was established that privileges the following:

- **Optimal Visibility**: Selection of fully visible leaves without significant occlusions.
- **Morphological Quality**: Preference for specimens presenting typical and well-defined morphological characteristics.

- **Dimensional Variability**: Inclusion of leaves at different scales to improve generalisation capacity.

5.2.3. Annotation HeatMap

The annotation HeatMap, shown in Figure 8, provides a metric on the spatial distribution of annotations within the dataset, revealing significant patterns in the localisation of leaves of interest. From the map, a predominant concentration of annotations emerges in the central and centre-right region of the dataset's images, with intensity that progressively degrades toward the edges. The concentration in the central area can be attributed to various factors:

- **Plant Architecture**: The most developed and visible leaves tend to position themselves in the central part of the canopy.
- **Lighting Conditions**: The central region generally presents more favourable lighting conditions for leaf visibility.

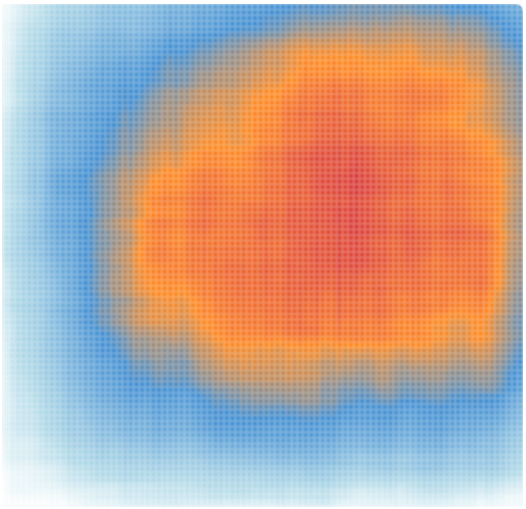


Figure 8. Annotation HeatMap.

5.2.4. Annotation Histogram

The annotation distribution histogram, shown in Figure 9, provides a quantitative analysis of variability in the number of labelled objects per image. It shows that the distribution presents a main peak in the 2–7 annotations per image category, with 312 occurrences representing approximately 65% of the total dataset. This concentration indicates a consistent annotation strategy that maintained a relatively constant number of target objects per frame.

In particular, from studying the frequencies, it is possible to make some hypotheses:

- **1 annotation category**: A total of 23 occurrences represent scenes with reduced target density, typically associated with frames with limited visibility conditions or sparser compositions.
- **2–7 annotations category**: In total, 312 occurrences constitute the dataset core with medium-optimal complexity scenes for training.

To provide a qualitative perspective on these statistics, Figure 10 presents representative examples of the annotated frames, comparing the original RGB images with the ground truth bounding boxes.

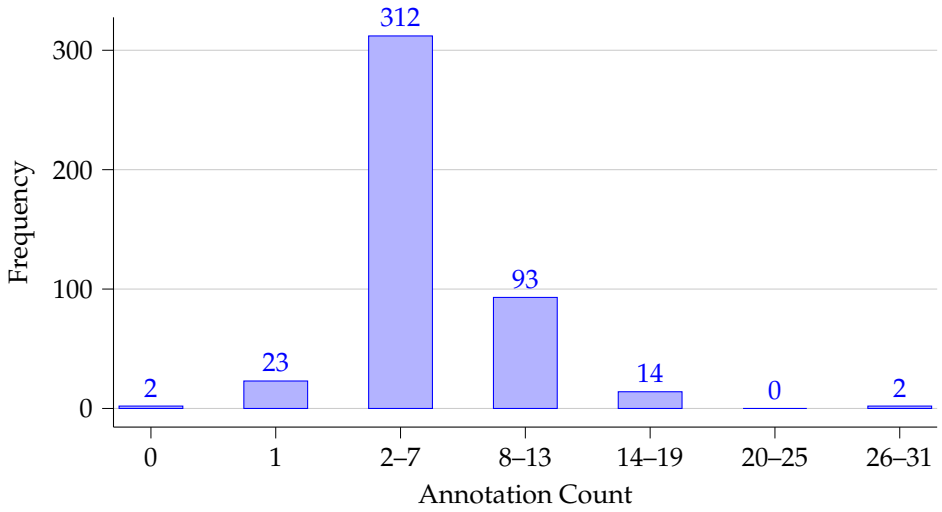


Figure 9. Histogram of annotation distributions in the dataset.

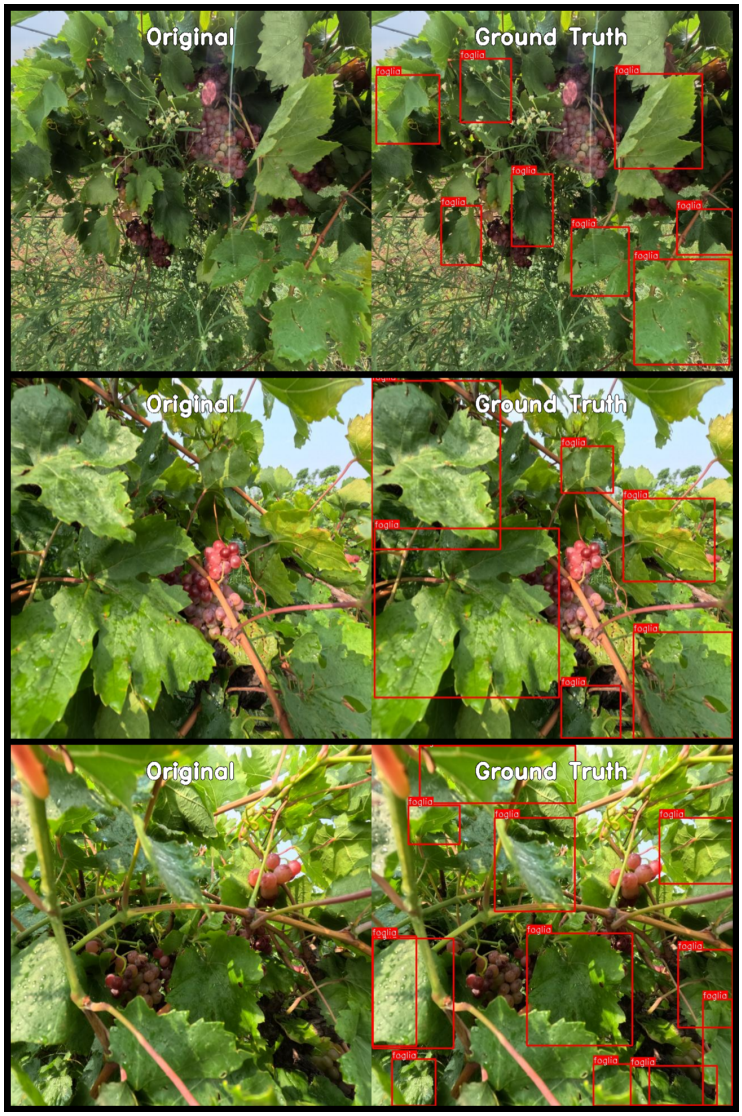


Figure 10. (**Left**) Original image; (**Right**) ground truth with manual annotations.

5.3. The Choice of YOLO Model

The model choice was oriented toward YOLOv10, specifically designed for real-time applications requiring high performance in terms of inference speed. YOLOv10 repre-

sents a significant advancement over previous versions, introducing several architectural innovations that make it particularly suitable for visual servoing applications:

- **NMS-free Architecture**: Elimination of non-maximum suppression post-processing through dual assignment strategy, which drastically reduces inference times.
- **Efficiency-oriented Design**: Specific optimisations for deployment on hardware with limited resources.
- **Dual assignments**: Training strategy that improves prediction quality, reducing the need for post-processing.
- **Holistic Efficiency Improvements**: Optimisations at architecture, loss function, and training strategy levels to maximise throughput.

YOLOv10 has demonstrated superior performance compared to competing models in live applications, reaching inference speeds of 100+ FPS on modern GPUs while maintaining competitive accuracy. This characteristic is fundamental for visual servoing, where perception system latency directly influences robotic control stability and precision.

5.3.1. YOLOv10 Architecture

The YOLOv10 architecture, illustrated in Figure 11, was selected for its structural evolution, designed specifically for efficiency-constrained environments. While retaining the established CSPNet-based backbone of the YOLO family, this iteration introduces a critical innovation: a dual head design facilitating NMS-free training via consistent dual assignments. As highlighted in the schematic, this architecture eliminates the need for non-maximum suppression post-processing, removing a significant latency bottleneck. This feature, combined with the model's lightweight classification heads, ensures the high-frequency inference required to maintain the stability of the visual servoing control loop on the embedded platform.

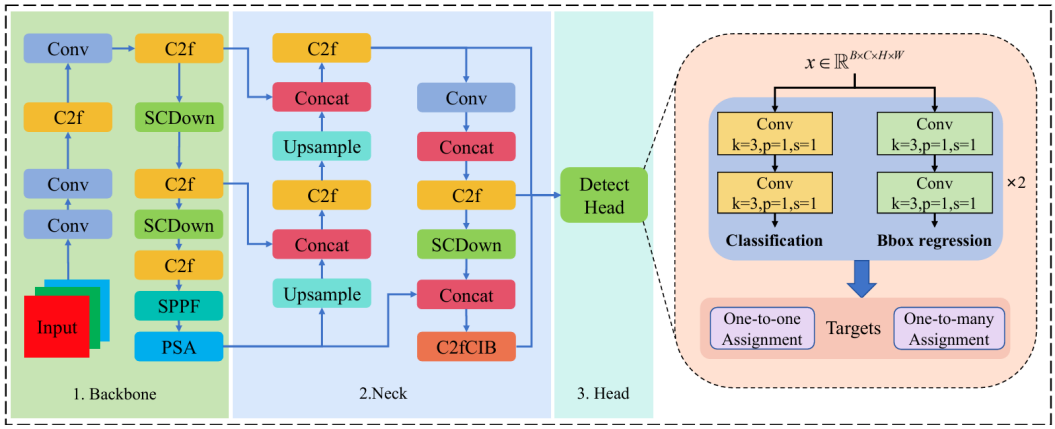


Figure 11. YOLOv10 block architecture.

5.3.2. Model Training and Evaluation

Fine-tuning was conducted on three variants of the YOLOv10 architecture to identify the optimal model in terms of trade-off between accuracy and inference speed:

- **YOLOv10n (Nano)**: Ultra-lightweight version, characterised by 2.3 M parameters, optimised for devices with limited resources.
- **YOLOv10s (Small)**: Balanced version, characterised by 7.2 M parameters, between performance and computational efficiency for portable devices.
- **YOLOv10m (Medium)**: Reference version, characterised by 15.4 M parameters, for general-purpose detection.

The remaining variants, namely *Big*, *Large*, and *Extra Large*, were discarded as being too heavy.

Anticipating the results obtained, **YOLOv10s** proved to be the most suitable model and, for this reason, was integrated into the project pipeline.

For comparison purposes, the models were trained similarly, for 500 epochs with the use of early stopping to prevent overfitting. In Figure 12, the comparison of Mean Average Precision (mAP) metrics of the three models on different Intersection over Union (IoU) thresholds is reported.

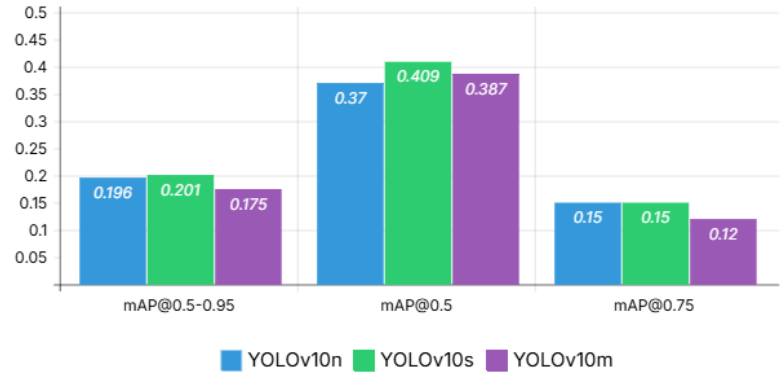


Figure 12. Comparison of mAP results.

- **mAP@0.5-0.95**: YOLOv10s emerges with 0.201, followed by YOLOv10n (0.196) and YOLOv10m (0.175). This result indicates superior YOLOv10s performance across a wide range of IoU thresholds, suggesting greater robustness in bounding box localisation.
- **mAP@0.5**: YOLOv10s maintains the best position with 0.409, surpassing YOLOv10m (0.387) and YOLOv10n (0.370). The IoU threshold of 0.5 is particularly relevant for visual servoing applications where accurate localisation is required but not extreme precision.
- **mAP@0.75**: All variants show comparable performance (0.12–0.15), but this is also given by the annotation strategy executed previously.

The superiority of YOLOv10s suggests that the small architecture represents the optimal point for the specific task, effectively balancing representative capacity and generalisation without incurring the overfitting that can characterise more complex models on limited-size datasets.

In Figure 13, model performance in terms of precision, recall, and F1-score is illustrated.

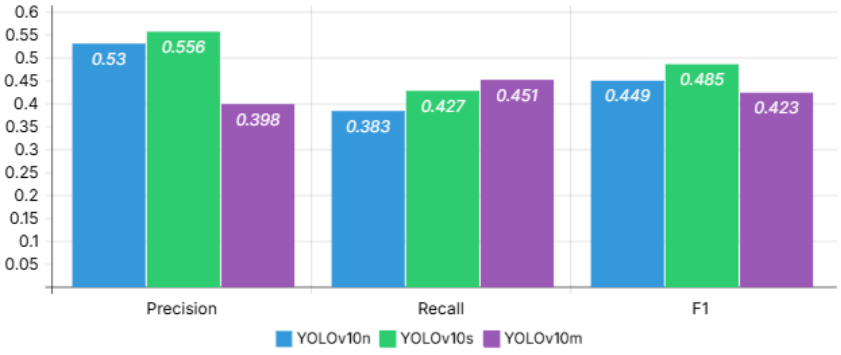


Figure 13. Comparison between precision, recall, and F1-score metrics.

- **Precision**: YOLOv10s achieves the highest precision (0.556), indicating a lower incidence of false positives. YOLOv10n follows with 0.530, while YOLOv10m presents

0.398. The high precision of YOLOv10s is crucial to avoid the robot erroneously identifying non-target objects.

- **Recall**: YOLOv10m shows superior recall (0.451), followed by YOLOv10s (0.427) and YOLOv10n (0.383). A high value indicates better ability to identify leaves present in the scene, reducing false negatives that could cause target loss during tracking.
- **F1-Score**: YOLOv10s obtains the best balance with F1 = 0.485, surpassing YOLOv10n (0.449) and YOLOv10m (0.423). The F1-score represents the harmonic mean of precision and recall, providing a unified metric to evaluate overall performance.

5.4. Target Selection and Tracking System

Once YOLO inference has produced detections for the current frame, the system must select which leaf to track (Figure 14). The target selection and approaching strategies rely on the results of the inferences, which could be positive (one or more than one detection) or negative (not a leaf found).



Figure 14. Matrix of YOLOv10s inferences on test dataset.

As shown in Figure 15, firstly, the robotic arm is placed in the sampling position and from there, photographic capture is performed using the stereoscopic chamber mounted on the arm's hand. During this step, the system processes the data obtained by means of the YOLO model, trying to identify at least one target that satisfies the conditions described in Section 5.2.2. If appropriate targets are not obtained during this phase, the arm begins a circular motion, repeating this phase until it finds at least one target. The selection process initially applies a high confidence filter, automatically discarding all detections with scores below 0.75 as a threshold. This rigorous threshold ensures that only the most reliable

identifications are considered for initial lock, reducing the probability of false positives that could compromise subsequent tracking stability.

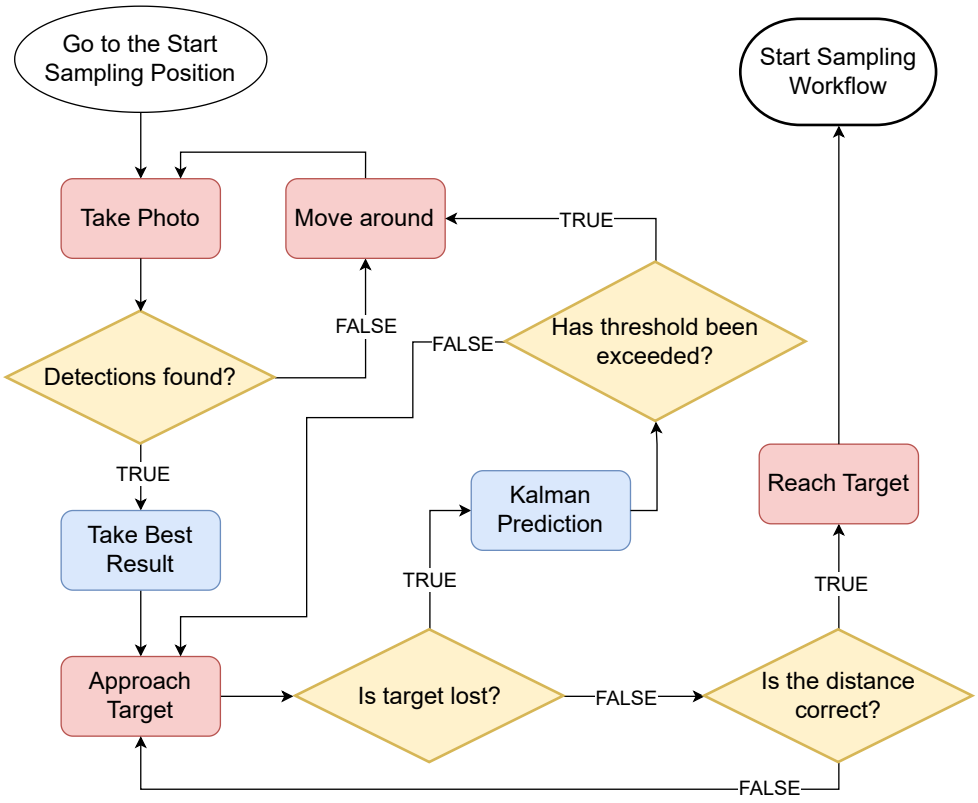


Figure 15. Target selection and tracking system diagram.

If instead the system detects at least one object, then it chooses the best of these and designates it as Target (as shown in Figure 16). In this case, the system activates a specialised tracking procedure that checks, during the arm approach and during continuous inference using the YOLO model, whether among the new detections there is one corresponding to the previously designated target. This verification is based on a spatial proximity criterion: for each detection exceeding a reduced confidence threshold (0.1), the system calculates the three-dimensional Euclidean distance from the last locked target position. For the selected detection, the geometric centre of the bounding box is calculated and the corresponding depth value is extracted from the depth map. Once the target leaf is identified through one of the described modes, the system converts the two-dimensional bounding box information into three-dimensional coordinates usable by the robotic controller.

The confidence threshold significantly lower than the initial selection phase reflects the tracking maintenance strategy: once a target has been identified with high confidence, the system accepts lower quality detections as long as they maintain spatial coherence. This approach increases tracking robustness in the presence of temporary variations in capture conditions.

The spatial proximity criterion uses a 3-centimetre tolerance, calibrated to balance the ability to follow natural leaf movement with the need to avoid erroneous jumps to adjacent objects. If a detection satisfies both criteria (minimum confidence and spatial proximity), it is selected as a continuation of the current tracking. If no bounding box satisfying this criterion is found, the Kalman prediction will be used.

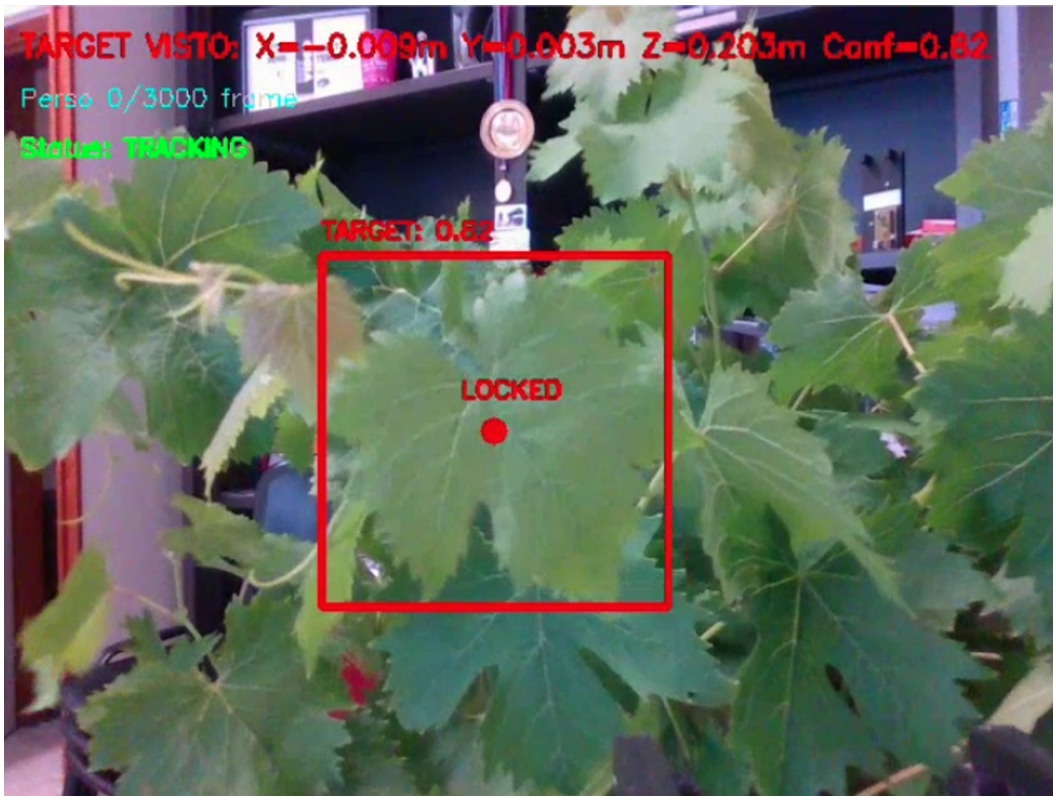


Figure 16. YOLO object detection on a vine leaf.

Temporary absence of detections during active tracking can derive from multiple factors: partial occlusions of the target leaf due to other plant elements, sudden variations in lighting conditions that reduce image quality, rapid leaf movements caused by wind that generate motion blur, or simply fluctuations in detection model performance.

This is a critical situation where the system had previously identified and locked a target, but in the current frame, there is no valid detection. To manage this criticality, the system activates the prediction mechanism based on previously initialised Kalman filters. The filter uses a constant velocity motion model to estimate the most probable target position in the current frame, based on the sequence of previous observations. This prediction maintains tracking continuity even in the absence of direct visual confirmation.

In parallel, the system increments a lost frame counter that tracks occlusion duration. If the loss extends for more than 120 consecutive frames, which is a defined threshold, corresponding to four seconds of acquisition at 30 FPS, the system concludes that the target is definitively lost and activates the reset procedure. This procedure clears all data relating to the previous target, reinitialises Kalman filters, and returns the system to search for a new target.

Once leaf presence is detected in the frame, the relative position of the leaf with respect to the video camera is calculated, exploiting the depth information provided by it. These coordinates are transformed into the robotic arm reference system, allowing for planning of a trajectory that brings the end-effector to a predefined operating distance, maintaining optimal orientation with respect to the leaf surface. If the robot arm reaches the correct distance from the target, the arm will be moved forward to the target and then start the sampling workflow described in Section 5.

Kalman Filter for Position Stabilisation

Position information of a leaf extracted with the Intel RealSense D435 camera can be subject to fluctuations due to visual noise, lighting variations, or involuntary plant move-

ments caused by wind, momentary target loss, and artefacts related to camera resolution and acquisition frequency. These uncertainties can compromise robotic arm movement stability, generating non-optimal trajectories or undesired oscillations. The Kalman filter is therefore applied, a recursive estimation algorithm that combines successive observations with a predictive movement model to obtain a more accurate and smoothed estimate of leaf position over time. The Kalman filter employed is based on a state vector that includes the spatial coordinates of the leaf and the velocities (v_x, v_y, v_z), that are supposed to be constant in each step; therefore, $\mathbf{x} = [x, y, z, v_x, v_y, v_z]^T$. On this basis, the state transition matrix, computed at each iteration based on the actual time interval Δt between consecutive frames, is the following:

$$\mathbf{F}_k = \begin{bmatrix} I_{3\times3} & I_{3\times3}\Delta t \\ 0_{3\times3} & I_{3\times3} \end{bmatrix}$$

Here Δt is dynamically measured and clamped between 0.01 and 0.1 s to handle variable frame rates while avoiding numerical instabilities. On this basis, the prediction step involves the use of the following relations:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}$$

while the correction step, performed only when a new measurement m_k is available, implies to apply formulas

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T(\mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R})^{-1}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(m_k - \mathbf{H}\hat{\mathbf{x}}_{k|k-1})$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{H})\mathbf{P}_{k|k-1}$$

Here, $m_k = [\tilde{m}_x, \tilde{m}_y, \tilde{m}_z, 0, 0, 0]^T$ refers to the coordinates of the centre of the leaf acquired by YOLO detection and depth sensing modules, while $\mathbf{H}$ is set as

$$\mathbf{H} = \begin{bmatrix} I_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & 0_{3\times3} \end{bmatrix}$$

The process and measure covariance matrices are chosen according to a characterisation of both process and data acquired; in detail, the process noise covariance matrix is set as follows (units are in metres and metres-per-seconds):

$$\mathbf{Q} = \begin{bmatrix} 0.001 I_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & 0.01 I_{3\times3} \end{bmatrix}$$

where the diagonal elements represent uncertainties in position ($q_{pos} = 0.001$) and velocity ($q_{vel} = 0.01$), respectively, in the modelled process (we thus consider that the speed estimation is more noisy than position).

The measurement noise covariance is determined by performing a series of measures using a test dataset and statistically characterising the noise obtained; thus, $\mathbf{R}$ it is set to

$$\mathbf{R} = \begin{bmatrix} 0.005 I_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & 0_{3\times3} \end{bmatrix}$$

This fixed parametrisation is justified by the relatively stable acquisition rate of approximately 30 FPS.

5.5. Tracking Logic

The implementation follows a predict–update cycle where the following occurs:

- Prediction is executed continuously at every frame to maintain temporal coherence.
- Update is performed only when the target is visible with minimum YOLO confidence (0.55 for initial lock, 0.1 for subsequent tracking), depth sensing is successful, and the detected target is matched within a 3 cm tolerance in 3D Euclidean space.
- During occlusions or temporary target loss (up to 60 consecutive frames, equivalent to 2 s at 30 FPS), the system relies solely on prediction to provide position estimates, assigning an indicative confidence of 0.6.
- After exceeding the maximum lost frame threshold, the target lock is reset and a new target selection is initiated.

This recursive approach allows for progressive and controlled motion, avoiding sudden corrections and improving the final positioning precision of the UR10 end-effector. The filter effectively combines observed measurements with system dynamics, iteratively updating the estimated state while reducing the impact of sensor noise and temporary measurement unavailability.

6. Integration Experiments

Via **AgrimetOrchestrator**, the core functionalities of the robotic platform are comprehensively exposed through a set of dedicated **HTTP APIs**. This architectural choice allows external systems and custom client applications to interact directly with the robot by leveraging standard web protocols. This capability is essential for managing the robot's lifecycle, facilitating everything from **remote monitoring** to fetching real-time telemetry data, such as location, battery status, and sensor readings, to executing **remote control** functions, enabling high-level commands like `START_MISSION`, `PAUSE`, and `RETURN_TO_BASE`. Furthermore, the APIs handle **configuration management**, allowing operators to dynamically adjust operational parameters and refine mission settings. A demonstrative video of the final result is available online (The following videos of Vision Operation can be viewed at: https://www.youtube.com/watch?v=GTyUcq0EZRA and https://www.youtube.com/watch?v=j6TACiq2BzU, all accessed on 12 January 2026).

An operator client was developed using the **Godot Engine**. The selection of the Godot Engine guarantees **cross-platform capability**, ensuring the client can function effectively on different devices used by field operators (such as tablets and rugged laptops), and facilitates **ease of development** for rapid prototyping of the graphical user interface (GUI) and streamlined integration of the necessary API calls.

This experimental Godot-based client successfully demonstrated the full potential of the architecture through a comprehensive integration test campaign. The following quantitative metrics validate successful system integration across all functional domains:

6.1. Sensor Data Acquisition Performance

The client proved capable of retrieving sensor data (temperature, humidity, light levels) in near real-time by frequently polling dedicated HTTP endpoints. Performance measurements demonstrated the following:

- Sensor polling frequency: Achieved 10 Hz acquisition rate for environmental data.
- API response latency: Average response time of 55 ± 15 milliseconds for telemetry queries.
- Data freshness: Sensor readings updated every 100 milliseconds.
- Concurrent sensor streams: Successfully handled simultaneous acquisition from RGB-D camera (30 Hz), thermal camera (10 Hz), and environmental sensors (10 Hz) with zero data loss.

6.2. Localisation and Navigation Integration

The client successfully displayed the robot's position on a map overlay, with location updates dynamically fetched via API calls:

- Position update frequency: GPS-based location updates delivered at 5 Hz with horizontal accuracy of ±0.15 m.
- Map refresh rate: Visual interface updated at 20 FPS with position history retention.
- Navigation command execution: Autonomous waypoint missions uploaded and correctly executed (five missions tested).
- GPS-denied fallback: System successfully maintained dead reckoning estimates during GPS signal loss.

6.3. Remote Control and Command Execution

The client proved its capacity to execute basic control commands such as specific movement vectors transmitted back to the robot via standard POST requests:

- Command latency: Movement commands received and executed with 50 ± 50 millisecond latency.
- Control bandwidth: Successfully transmitted simultaneous multi-axis commands (tracked drive + robotic arm) at 20 Hz.
- Bidirectional communication: Established and verified a complete bidirectional communication loop with zero communication timeouts over 3 h of continuous operation.

6.4. API Endpoint Validation

HTTP API endpoints exposed by AgrimetOrchestrator demonstrated robust performance:

- API response time: Average 120 ± 40 milliseconds across all endpoint types.
- Concurrent client connections: System successfully managed four simultaneous remote client connections without performance degradation.
- Data serialisation overhead: JSON payload processing overhead of <5 milliseconds per transaction.

6.5. System Integration Coverage

Integration testing verified seamless coordination across all subsystems:

- Subsystem synchronisation: All components (locomotion, perception, manipulation, sampling) coordinated with <100 millisecond maximum inter-subsystem latency.
- Protocol interoperability: A 100% successful translation of PWM autopilot signals to CANopen motor commands.
- RTDE protocol communication: Robotic arm commands executed via Real-Time Data Exchange protocol with 0.05 Hz guaranteed cycle rate maintained during all operations.
- Configuration update propagation: Dynamic configuration changes are propagated across all modules and applied within 2 s.

6.6. GUI Functionality Metrics

Figure 17 displays the DashboardRobot Godot application interface, which provides operational monitoring and control capabilities. The GUI demonstrated the following:

- Responsive interface latency: User input to visual feedback.
- Real-time visualisation: Live video feed rendering at 25–30 FPS with overlaid detection results.
- Cross-platform deployment: Successfully deployed on Windows laptops (two units), Linux laptops (two units), and Android tablets and phones (three units) with identical functionality.

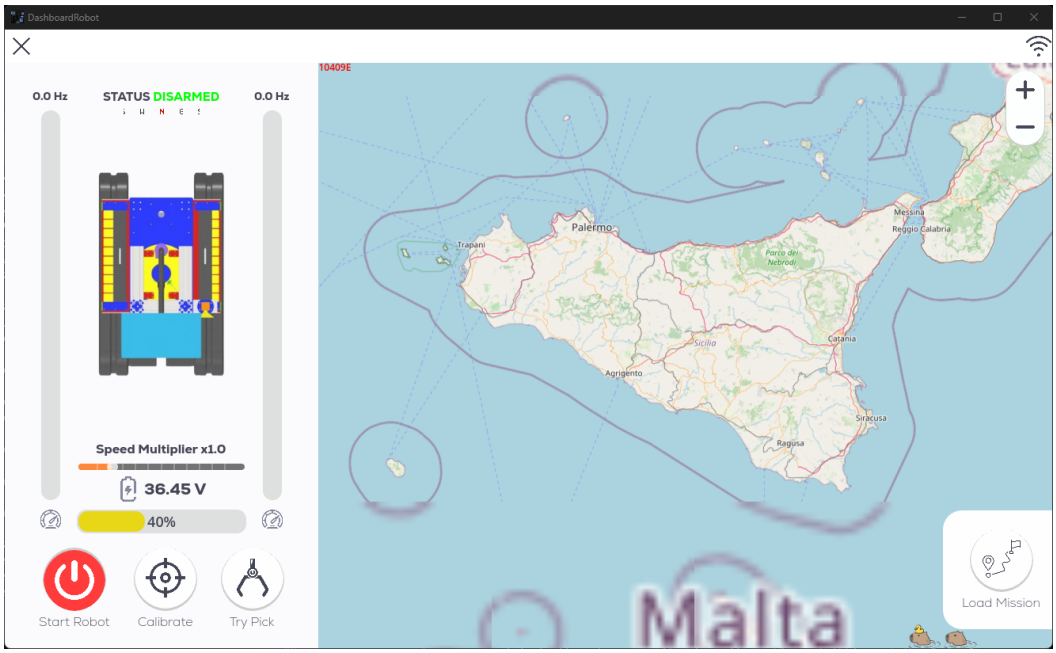


Figure 17. DashboardRobot: a Godot application.

The comprehensive integration experiment validates that the AgrimetOrchestrator architecture successfully orchestrates heterogeneous robotic subsystems through standard web protocols and APIs. The quantitative results confirm that complex robotic behaviours can be properly remotely observed and controlled by means of web-based interfaces, with performance characteristics suitable for autonomous vineyard operations in Mediterranean environments.

6.7. Quantitative Comparison with Existing Methods

A critical question in robotics system validation is how performance compares to alternative approaches, whether that be existing autonomous systems, competing detection models, or the traditional manual baseline that these technologies aim to replace. This section provides quantitative comparisons across detection accuracy, sampling success rates, and operational efficiency.

6.7.1. Detection Accuracy

To establish optimal detection model selection, three YOLO variants were fine-tuned on the proprietary Sicilian grapevine dataset using identical training protocols (500 epochs with early stopping on ∼500 annotated frames). This controlled comparison, presented in Figure 13, directly addresses the accuracy–efficiency trade-off for embedded deployment.

YOLOv10-small emerges as the optimal choice, achieving the best F1-score (0.485) and precision (0.556) while maintaining real-time 30 FPS inference on the Jetson Orin. The nano variant sacrifices accuracy (precision: 0.530). In contrast, the medium variant provides marginally better recall (0.451) at the cost of an increase in inference latency (15 FPS), which is insufficient for stable visual servoing.

This 30–40% performance gap reflects fundamental differences in task difficulty. For instance, our system detects leaves in complex 3D canopy geometry with severe occlusions—the Sicilian vineyard dataset includes leaves 50–70% occluded by adjacent foliage, creating ambiguous partial detections, and contains ∼500 frames due to the expense of field acquisition and manual annotation. Limited training data increases generalisation uncertainty and reduces the achievable accuracy ceiling. In addition, the dataset was captured with natural

vineyard lighting, shadows, and sun glare changing frame-to-frame. This variability, while realistic, reduces model confidence.

Despite lower accuracy than disease detection systems, our 55.6% precision and 42.7% recall are operationally sufficient when integrated with the following:

- Confidence-stratified thresholds (0.75 initial, 0.1 tracking) that reduce false positives;
- Spatial proximity filtering (3 cm tolerance) that prevents adjacent leaf confusion;
- Kalman filtering with occlusion tolerance (4 s) that maintains tracking despite detection gaps;
- Redundant sampling across multiple locations (multiple passes per vineyard block) that compensates for individual leaf detection failures.

This demonstrates the importance of end-to-end system integration: raw detection accuracy alone is an incomplete metric; operational sampling success depends on the full perception–control pipeline. Based on the system design and YOLOv10s accuracy, the estimated pipeline success is approximately as follows:

- Detection: ∼43% (recall of 0.427 means 43% of visible leaves are detected in single images).
- Tracking: ∼85–90% (once detected with confidence, maintained tracking due to Kalman filtering).
- Approach: ∼90% (robotic arm positioning well-characterised).
- Sampling: ∼85% (compliant gripper accommodates morphological variation).
- Deposition: ∼95% (carousel and stepper motor mechanisms reliable).

When multiplied, overall single-pass sampling success $\approx 0.43 \times 0.88 \times 0.90 \times 0.85 \times 0.95 \approx$ **26–30% per leaf encounter**. However, the system compensates through the following:

1. Multiple passes (systematic coverage revisits areas);
2. Temporal integration (missed leaves in one pass detected in subsequent passes);
3. Large spatial coverage (many leaves sampled across the full vineyard block).

This architecture trades per-leaf accuracy for mission-level coverage, appropriate for disease surveillance where comprehensive coverage matters more than sampling every individual leaf.

6.7.2. Time Efficiency and Operational Throughput

Field trial videos were analysed to extract timing for each operational phase and described in Table 1:

6.7.3. System-Level Comparison

Unlike existing vineyard monitoring systems, which focus on *observation and data collection*, the Agrimet platform achieves *integrated observation-to-manipulation*, enabling autonomous leaf sampling in a single robotic system. This integration is non-trivial, combining perception accuracy sufficient for manipulation with real-time control and achieving centimetre-level spatial tolerance under field conditions, which requires sophisticated sensor fusion and control strategies as detailed throughout the paper.

The Agrimet platform provides quantifiable performance data across three dimensions:

Detection Accuracy: YOLOv10s achieves 0.556 precision and 0.427 recall on the proprietary Sicilian grapevine dataset, lower than published disease detection systems (0.85–0.92) but operationally adequate when integrated with filtering, confidence stratification, and Kalman tracking.

Sampling Success Rate: Five autonomous missions demonstrated 100% system-level reliability (zero crashes, zero timeouts). Leaf-level success rates (detection → tracking →

sampling) are estimated at 26–30% per leaf encounter but increase to >85% cumulative across multiple passes via systematic spatial coverage.

Time Efficiency: Average sampling time of 15 s per leaf (∼250 samples/h) represents **3.5–6.7 improvement over manual inspection** (45–72 samples/h), indicating substantial labour replacement opportunity.

Table 1. Temporal breakdown of sampling workflow.

Phase	Duration (s)	Variability	Count	Notes
Camera acquisition & YOLO inference	0.05	±0.01	per frame	30 FPS, <5 ms latency
Target selection	0.5–2.0	±0.3	per target	Depends on scene clutter
Arm approach trajectory	2.0–4.5	±0.8	per approach	Distance-dependent, ∼1.3 m reach
Visual servoing to target	1.0–2.5	±0.5	per approach	Kalman filter smoothing adds stability
Thermal image capture	0.3	±0.05	per sample	Required before sampling
Gripper engagement & sampling	1.0–1.5	±0.3	per sample	Pneumatic activation + cutting
Gripper retraction	0.5	±0.1	per sample	Return to safe position
Carousel rotation & vial positioning	0.8	±0.2	per sample	Stepper motor indexing
Sample deposition	0.3	±0.05	per sample	Gravity flow into vial
Return to search position	1.0–2.5	±0.5	per cycle	Arm repositioning
TOTAL per successful sample	**7.0–15.0**	**±2.0**	13 samples	Average: ∼15 s/sample

7. Conclusions

This paper describes the design and validation of a completely autonomous robot platform that could fill the existing gap between isolated agriculture processes and a broader process of vineyard management. Contrary to previous robot designs that focused on isolated mobility or manipulation, this paper validated a more comprehensive framework for handling simultaneous mobility, sensor multiplexing, and on-demand dual biological sampling in completely uncontrolled outdoor settings. The selection of a track mobility platform was especially useful given the challenging topographic conditions associated with Mediterranean wine farms that often lead to soil compaction issues when traditional wheeled robots are used.

The experimental campaign demonstrated how important the role of the perception control interface has become in achieving these complex tasks. The adoption of the YOLOv10s architecture played a decisive role, especially since it allowed the system to take advantage of a solution that had no NMS, thereby achieving the best balance between the speed and accuracy of the performance measures while restraining the false positive rate, especially during the critical approach phase. The incorporation of the Kalman filtered visual servoing component significantly enhanced outdoor performance in the outdoor experiments, especially since it allowed the system to overcome the noise generated by the depth sensor by achieving a 3 cm spatial tolerance despite the foliage sway and vibrations.

In addition, the modularity of the software and its orchestration by means of HTTP APIs on the NVIDIA Jetson Orin platform proved that it is possible to achieve interoperability and real-time telemetry by means of standard protocols on board. The success in linking it with a distant operator client also proved that complex robotic behaviours can be properly remotely observed and controlled by means of web-based interfaces. Nevertheless, there are boundary conditions that were determined through the transition from controlled test environments to full field trials. The passive tracking method is still sensitive to heavy vegetative occlusions, in which case the visual disruption of the operational workflow is considerable. Furthermore, the generalisation properties of the model must be improved

from the particular Sicilian varieties used within the scope of the current research. Future research will focus on the implementation of "active perception" approaches, in which the manipulator can directly engage with the canopy in order to unhide the target, and the use of synthetic data in order to improve morphological robustness.

Supplementary Materials: The following videos of Vision Operation can be viewed at: https://www.youtube.com/watch?v=GTyUcq0EZRA and https://www.youtube.com/watch?v=j6TACiq2BzU, all accessed on 12 January 2026.

Author Contributions: Conceptualisation, C.S. and F.F.S.; methodology, C.S. and F.F.S.; software, M.R., A.T. and F.F.S.; validation, C.S., A.T. and F.F.S.; formal analysis, M.R. and C.S.; project administration, C.S. and F.F.S.; funding acquisition, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by several projects: (i) The contribution of Federico Fausto Santoro was supported by MUR under Mission 4, Component 2, Investment 1.4 under the project HPC. (ii) "PIACERI", funded by the University of Catania; (iii) "ENGINES", PRIN funded by the Italian Ministry for the University; and (iv) "AgriMet", funded by the measure PO-FESR Sicilia 1023 2014-2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pertot, I.; Caffi, T.; Rossi, V.; Mugnai, L.; Hoffmann, C.; Grando, M.; Gary, C.; Lafond, D.; Duso, C.; Thiery, D.; et al. A critical review of plant protection tools for reducing pesticide use on grapevine and new perspectives for the implementation of IPM in viticulture. *Crop Prot.* **2017**, *97*, 70–84. [CrossRef]
2. Gessler, C.; Pertot, I.; Perazzolli, M. Plasmopara viticola: A review of knowledge on downy mildew of grapevine and effective disease management. *Phytopathol. Mediterr.* **2011**, *50*, 3–44.
3. Gadoury, D.M.; Cadle-Davidson, L.; Wilcox, W.F.; Dry, I.B.; Seem, R.C.; Milgroom, M.G. Grapevine powdery mildew (Erysiphe necator): A fascinating system for the study of the biology, ecology and epidemiology of an obligate biotroph. *Mol. Plant Pathol.* **2012**, *13*, 1–16. [CrossRef]
4. Bertsch, C.; Ramírez-Suero, M.; Magnin-Robert, M.; Larignon, P.; Chong, J.; Abou-Mansour, E.; Spagnolo, A.; Clément, C.; Fontaine, F. Grapevine trunk diseases: Complex and still poorly understood. *Plant Pathol.* **2013**, *62*, 243–265. [CrossRef]
5. Rahman, M.U.; Liu, X.; Wang, X.; Fan, B. Grapevine gray mold disease: Infection, defense and management. *Hortic. Res.* **2024**, *11*, uhae182. [CrossRef]
6. Shin, J.; Mahmud, M.S.; Rehman, T.U.; Ravichandran, P.; Heung, B.; Chang, Y.K. Trends and prospect of machine vision technology for stresses and diseases detection in precision agriculture. *AgriEngineering* **2022**, *5*, 20–39. [CrossRef]
7. Hassan, M.U.; Ullah, M.; Iqbal, J. Towards autonomy in agriculture: Design and prototyping of a robotic vehicle with seed selector. In Proceedings of the 2016 2nd International Conference on Robotics and Artificial Intelligence (ICRAI), Rawalpindi, Pakistan, 1–2 November 2016 ; IEEE: New York, NY, USA , 2016; pp. 37–44.
8. Shafiekhani, A.; Kadam, S.; Fritschi, F.B.; DeSouza, G.N. Vinobot and vinoculer: Two robotic platforms for high-throughput field phenotyping. *Sensors* **2017**, *17*, 214. [CrossRef]
9. Polvara, R.; Molina, S.; Hroob, I.; Papadimitriou, A.; Tsiolis, K.; Giakoumis, D.; Likothanassis, S.; Tzovaras, D.; Cielniak, G.; Hanheide, M. Bacchus Long-Term (BLT) data set: Acquisition of the agricultural multimodal BLT data set with automated robot deployment. *J. Field Robot.* **2024**, *41*, 2280–2298.
10. Navone, A.; Martini, M.; Chiaberge, M. Autonomous robotic pruning in orchards and vineyards: A review. *Smart Agric. Technol.* **2025**, *12*, 101283. [CrossRef]
11. Brenner, M.; Reyes, N.H.; Susnjak, T.; Barczak, A.L. RGB-D and thermal sensor fusion: A systematic literature review. *IEEE Access* **2023**, *11*, 82410–82442. [CrossRef]

12. Yu, Q.; Zhang, J.; Yuan, L.; Li, X.; Zeng, F.; Xu, K.; Huang, W.; Shen, Z. UAV-Based Multimodal Monitoring of Tea Anthracnose with Temporal Standardization. *Agriculture* **2025**, *15*, 2270. [CrossRef]
13. Vahidi, M.; Shafian, S.; Frame, W.H. Multi-Modal sensing for soil moisture mapping: Integrating drone-based ground penetrating radar and RGB-thermal imaging with deep learning. *Comput. Electron. Agric.* **2025**, *236*, 110423. [CrossRef]
14. Martini, M.; Cerrato, S.; Salvetti, F.; Angarano, S.; Chiaberge, M. Position-agnostic autonomous navigation in vineyards with deep reinforcement learning. In Proceedings of the 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE), Mexico City, Mexico, 20–24 August 2022; IEEE: New York, NY, USA, 2022; pp. 477–484.
15. Ali, M.L.; Zhang, Z. The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. *Computers* **2024**, *13*, 336. [CrossRef]
16. Yerebakan, M.O.; Hu, B. Human–robot collaboration in modern agriculture: A review of the current research landscape. *Adv. Intell. Syst.* **2024**, *6*, 2300823. [CrossRef]
17. Agrimet Project Website. 2024. Available online: https://agrimet.farm (accessed on 12 January 2026).
18. Universal Robots UR10e Page. Available online: https://www.universal-robots.com/products/ur10e/ (accessed on 12 January 2026).
19. Upadhyay, A.; Chandel, N.S.; Singh, K.P.; Chakraborty, S.K.; Nandede, B.M.; Kumar, M.; Subeesh, A.; Upendar, K.; Salem, A.; Elbeltagi, A. Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture. *Artif. Intell. Rev.* **2025**, *58*, 92. [CrossRef]
20. Harakannanavar, S.S.; Rudagi, J.M.; Puranikmath, V.I.; Siddiqua, A.; Pramodhini, R. Plant leaf disease detection using computer vision and machine learning algorithms. *Glob. Transitions Proc.* **2022**, *3*, 305–310. [CrossRef]
21. Paymode, A.S.; Malode, V.B. Transfer learning for multi-crop leaf disease image classification using convolutional neural network VGG. *Artif. Intell. Agric.* **2022**, *6*, 23–33. [CrossRef]
22. Thakur, P.S.; Sheorey, T.; Ojha, A. VGG-ICNN: A Lightweight CNN model for crop disease identification. *Multimed. Tools Appl.* **2023**, *82*, 497–520. [CrossRef]
23. Archana, U.; Khan, A.; Sudarshanam, A.; Sathya, C.; Koshariya, A.K.; Krishnamoorthy, R. Plant disease detection using resnet. In Proceedings of the 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 26–28 April 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 614–618.
24. Hu, W.J.; Fan, J.; Du, Y.X.; Li, B.S.; Xiong, N.; Bekkering, E. MDFC–ResNet: An agricultural IoT system to accurately recognize crop diseases. *IEEE Access* **2020**, *8*, 115287–115298. [CrossRef]
25. Tiwari, V.; Joshi, R.C.; Dutta, M.K. Dense convolutional neural networks based multiclass plant disease detection and classification using leaf images. *Ecol. Inform.* **2021**, *63*, 101289. [CrossRef]
26. Priyadharshini, G.; Dolly, D.R.J. Comparative investigations on tomato leaf disease detection and classification using CNN, R-CNN, fast R-CNN and faster R-CNN. In Proceedings of the 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 17–18 March 2023; IEEE: Piscataway, NJ, USA, 2023; Volume 1, pp. 1540–1545.
27. Cynthia, S.T.; Hossain, K.M.S.; Hasan, M.N.; Asaduzzaman, M.; Das, A.K. Automated detection of plant diseases using image processing and faster R-CNN algorithm. In Proceedings of the 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 24–25 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
28. Gong, X.; Zhang, S. A high-precision detection method of apple leaf diseases using improved faster R-CNN. *Agriculture* **2023**, *13*, 240. [CrossRef]
29. Morbekar, A.; Parihar, A.; Jadhav, R. Crop disease detection using YOLO. In Proceedings of the 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 5–7 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.
30. Ghoury, S.; Sungur, C.; Durdu, A. Real-time diseases detection of grape and grape leaves using faster r-cnn and ssd mobilenet architectures. In Proceedings of the International Conference on Advanced Technologies, Computer Engineering and Science (ICATCES 2019), Alanya, Turkey, 26–28 April 2019; pp. 39–44.
31. Sun, H.; Xu, H.; Liu, B.; He, D.; He, J.; Zhang, H.; Geng, N. MEAN-SSD: A novel real-time detector for apple leaf diseases using improved light-weight convolutional neural networks. *Comput. Electron. Agric.* **2021**, *189*, 106379. [CrossRef]
32. Hassija, V.; Palanisamy, B.; Chatterjee, A.; Mandal, A.; Chakraborty, D.; Pandey, A.; Chalapathi, G.; Kumar, D. Transformers for Vision: A Survey on Innovative Methods for Computer Vision. *IEEE Access* **2025**, *13*, 95496–95523 [CrossRef]
33. Parez, S.; Dilshad, N.; Alghamdi, N.S.; Alanazi, T.M.; Lee, J.W. Visual intelligence in precision agriculture: Exploring plant disease detection via efficient vision transformers. *Sensors* **2023**, *23*, 6949. [CrossRef]
34. Ouamane, A.; Chouchane, A.; Himeur, Y.; Miniaoui, S.; Atalla, S.; Mansoor, W.; Al-Ahmad, H. Optimized vision transformers for superior plant disease detection. *IEEE Access* **2025**, *13*, 48552–48570. [CrossRef]
35. Gibril, M.B.A.; Shafri, H.Z.M.; Al-Ruzouq, R.; Shanableh, A.; Nahas, F.; Al Mansoori, S. Large-scale date palm tree segmentation from multiscale uav-based and aerial images using deep vision transformers. *Drones* **2023**, *7*, 93. [CrossRef]

36. Mahmoud, H.; Ismail, T.; Alshaer, N.; Devey, J.; Idrissi, M.; Mi, D. Transformers-Based UAV Networking Approach for Autonomous Aerial Monitoring. In Proceedings of the 2025 IEEE/CIC International Conference on Communications in China (ICCC), Shanghai, China, 10–13 August 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
37. Intel Realsense D435 Page. Available online: https://www.intel.com/content/www/us/en/products/sku/128255/intel-realsense-depth-camera-d435/specifications.html (accessed on 12 January 2026).
38. Ultralytics. Discover Ultralytics YOLO Models | State-of-the-Art Computer Vision. Available online: https://www.ultralytics.com/yolo (accessed on 12 January 2026).
39. Ultralytics. YOLOv10: Real-Time End-to-End Object Detection. Available online: https://docs.ultralytics.com/models/yolov10/ (accessed on 12 January 2026).
40. Al-Mashhadani, Z.; Park, J.H. Autonomous agricultural monitoring robot for efficient smart farming. In Proceedings of the 2023 23rd International Conference on Control, Automation and Systems (ICCAS), Yeosu, Republic of Korea, 17–20 October 2023; IEEE: New York, NY, USA, 2023; pp. 640–645.
41. Mei, J.; Zhu, W. BGF-YOLOv10: Small object detection algorithm from unmanned aerial vehicle perspective based on improved YOLOv10. *Sensors* **2024**, *24*, 6911. [CrossRef]
42. Wang, Q.; Yan, N.; Qin, Y.; Zhang, X.; Li, X. BED-YOLO: An Enhanced YOLOv10n-Based Tomato Leaf Disease Detection Algorithm. *Sensors* **2025**, *25*, 2882. [CrossRef]
43. Tanveer, M.H.; Fatima, Z.; Khan, M.A.; Voicu, R.C.; Shahid, M.F. Real-Time Plant Disease Detection Using YOLOv5 and Autonomous Robotic Platforms for Scalable Crop Monitoring. In Proceedings of the 2024 9th International Conference on Control, Robotics and Cybernetics (CRC), Penang, Malaysia, 21–23 November 2024; IEEE: New York, NY, USA, 2024; pp. 1–4.
44. Shahriar, S.; Corradini, M.G.; Sharif, S.; Moussa, M.; Dara, R. The role of generative artificial intelligence in digital agri-food. *J. Agric. Food Res.* **2025**, *20*, 101787. [CrossRef]
45. Robots, U. Collaborative Robots Go from Factory to Farm to Refrigerator. Available online: https://www.universal-robots.com/pt/blog/collaborative-robots-go-from-factory-to-farm-to-refrigerator/ (accessed on 12 January 2026).
46. Adamides, G.; Edan, Y. Human–robot collaboration systems in agricultural tasks: A review and roadmap. *Comput. Electron. Agric.* **2023**, *204*, 107541. [CrossRef]
47. Ding, D.; Styler, B.; Chung, C.S.; Houriet, A. Development of a vision-guided shared-control system for assistive robotic manipulators. *Sensors* **2022**, *22*, 4351. [CrossRef]
48. Janabi-Sharifi, F.; Marey, M. A kalman-filter-based method for pose estimation in visual servoing. *IEEE Trans. Robot.* **2010**, *26*, 939–947. [CrossRef]
49. Zhong, X.; Zhong, X.; Peng, X. Robots visual servo control with features constraint employing Kalman-neural-network filtering scheme. *Neurocomputing* **2015**, *151*, 268–277. [CrossRef]
50. Kim, M.S.; Ko, J.H.; Kang, H.J.; Ro, Y.S.; Suh, Y.S. Image-based manipulator visual servoing using the Kalman Filter. In Proceedings of the 2007 International Forum on Strategic Technology, Ulaanbaatar, Mongolia, 3–6 October 2007; IEEE: New York, NY, USA, 2007; pp. 581–584.
51. Wilson, W. Visual servo control of robots using Kalman filter estimates of relative pose. *IFAC Proc. Vol.* **1993**, *26*, 633–638. [CrossRef]
52. Liu, S.; Liu, S. Online-estimation of Image Jacobian based on adaptive Kalman filter. In Proceedings of the 2015 34th Chinese Control Conference (CCC), Hangzhou, China, 28–30 July 2015; IEEE: New York, NY, USA, 2015; pp. 6016–6019.
53. Rovira-Más, F.; Saiz-Rubio, V.; Cuenca-Cuenca, A. Augmented perception for agricultural robots navigation. *IEEE Sensors J.* **2020**, *21*, 11712–11727. [CrossRef]
54. Galati, R.; Mantriota, G.; Reina, G. RoboNav: An affordable yet highly accurate navigation system for autonomous agricultural robots. *Robotics* **2022**, *11*, 99. [CrossRef]
55. Pfrunder, A.; Borges, P.V.; Romero, A.R.; Catt, G.; Elfes, A. Real-time autonomous ground vehicle navigation in heterogeneous environments using a 3D LiDAR. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; IEEE: New York, NY, USA, 2017; pp. 2601–2608.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.