UNIVERSITÀ DEGLI STUDI DI TORINO

SCUOLA DI DOTTORATO

DOTTORATO IN
SCIENZE AGRARIE, FORESTALI E ALIMENTARI

CICLO: XXXVII

Geomatic Techniques to Support Phytosanitary Products Tests whithin the EPPO Standard Framework

Samuele Bumbaca

Docente guida:
Prof. Enrico Corrado Borgogno Mondino

Coordinatore del Ciclo:

Prof. Domenico Bosco

ANNI
2023; 2024; 2025

Abstract

Background and Research Gap: European agricultural research for Plant Protection Product (PPP) development relies on statistical hypothesis testing under European and Mediterranean Plant Protection Organization (EPPO) standards. Traditional statistical approaches recognized by EPPO attempt to address environmental variability through experimental design features such as randomized controls and blocking. These methods are fundamentally limited by their reliance on a priori identification of variance sources, where experimentalists must subjectively identify environmental variability patterns before data collection based solely on human experience and field observation. Geostatistical methods offer a mathematically rigorous alternative by enabling the estimation of spatial environmental variability after data collection through variograms or spline fitting techniques. While these methods are recognized by EPPO and have demonstrated superior performance in modeling environmental heterogeneity, they require spatially referenced observations (each data point must have precise spatial coordinates). This requirement creates a significant implementation barrier, as traditional manual assessment methods used in PPP trials typically do not capture spatial coordinates, making it practically impossible to apply geostatistical approaches despite their theoretical advantages. This research gap (the absence of practical methods to generate spatially referenced datasets that would enable geostatistical analysis within EPPO-compliant trials) has prevented the widespread adoption of more robust statistical approaches in agricultural field studies.

Research Objectives: This research addresses the identified gap by investigating the applicability of geomatics technologies for recording spatially referenced observations in compliance with EPPO standards. The objective is to establish practical methods for generating georeferenced datasets that enable geostatistical analysis, demonstrating how these techniques can facilitate the adoption of more robust statistical approaches in agricultural research within the EPPO standard framework.

Methodology: This work considered three aspects of geomatics technologies applicability in this context:

1. counting, using deep learning object detectors to count maize seedlings on orthomosaics;
2. scoring, using machine learning regressors to score phytotoxicity via photogrammetric multispectral imaging and custom feature extraction;
3. classify, using anomaly detection to classify healthy or deseased plant tissues via pre-trained models.

Key Results: To achieve EPPO benchmark performance (R^2 > 0.85) in maize seedling counting from orthomosaics, transformer-based object detectors required approximately 60 labeled training images (225×225 pixels at $5\,\mathrm{mm/pixel}$ resolution), with domain-specific training data proving essential for agricultural applications. Machine learning successfully automated phytotoxicity scoring (κ > 0.7) with only 30 training samples by combining photogrammetric 3D models with spectral imaging, converting subjective ordinal assessments into objective continuous measurements suitable for enhanced statistical analysis. Pre-trained models achieved

accurate plant deseases classification (accuracy > 0.85) using anomaly detection techniques without requiring task-specific training. All three approaches successfully demonstrated the feasibility of automatically collecting georeferenced observations for agricultural trial assessments that comply with EPPO standards, thereby enabling the implementation of geostatistical methods for improved spatial analysis.

Novel Contribution: This work establishes the first systematic evaluation of minimum requirements for implementing geomatics techniques within EPPO standards, providing practical guidelines for dataset size, validation protocols, and integration strategies. The research demonstrates that geomatics technologies can successfully provide spatial coordinates alongside observations for all EPPO variable types, enabling the implementation of geostatistical methods that improve environmental variability modeling compared to traditional experimental design approaches.

Practical Impact: The findings enable agricultural researchers to adopt more robust statistical methods for PPP trials by providing clear implementation requirements for digital data collection technologies, ultimately leading to more accurate efficacy evaluations and improved crop protection strategies.

Contents

Introduction

1.1 Research Overview and Motivation

Agricultural research for plant protection product (PPP) development relies heavily on statistical hypothesis testing to establish efficacy and safety. Traditional experimental approaches in this field face a fundamental limitation: they require experimentalists to identify and control for environmental variability before data collection through experimental design features like randomized controls and blocking strategies. This "a priori" identification of variance sources depends heavily on experimentalist expertise and field experience rather than systematic statistical procedures, creating potential biases and limitations in agricultural field studies.

The European and Mediterranean Plant Protection Organization (EPPO) has established comprehensive standards for PPP evaluation that require rigorous statistical validation across diverse data types. However, classic statistical frameworks within these standards struggle to adequately address unknown spatial environmental variability that emerges during

trials.

Geostatistical methods provide a well-established alternative approach, enabling the estimation of spatial environmental variability through mathematical techniques such as variograms and spline fitting. This approach fundamentally shifts the traditional paradigm by eliminating the need to have prior knowledge of environmental variability before starting the field trial. Instead, it enables this variability to be estimated afterwards, even when its effects overlap or interact with those of the treatments. While this analysis capability represents a significant advantage over classical experimental design approaches, geostatistics requires spatially referenced data collection, introducing the practical challenge of precise observation geolocation in agricultural field conditions.

Recent advances in geomatics technologies - including photogrammetry, spectral imaging, and machine learning - present new opportunities to overcome these limitations by providing efficient methods for generating spatially referenced datasets. However, the practical implementation of these technologies within the regulatory framework of PPP evaluation remains unexplored.

1.2 Research Objectives

This research investigates the applicability of geomatics technologies for recording spatially referenced observations across all EPPO standard categories. The overarching goal is to demonstrate how these technologies can facilitate the adoption of geostatistical methods in hypothesis testing for agricultural research conducted within the EPPO framework.

EPPO standard assessments were categorized into three main variable types: continuous or discrete, ordinal, and nominal or binary. One representative assessment was selected for each variable type:

1. Plant counting (continuous or discrete)
2. Phytotoxicity scoring (ordinal)
3. Disease detection (nominal or binary)

Each of these assessments was addressed through the application of geomatics technologies to demonstrate the feasibility of collecting georeferenced observations with precision and accuracy that meet EPPO standard requirements.

1.3 Thesis Structure

The structure of the thesis is as follows:

Theoretical Background and Methodology presents the theoretical foundation, including an overview of Plant Protection Product (PPP) regulations, relevant EPPO standards, and the methodological background covering geostatistics, photogrammetry, spectral imaging, and machine learning techniques.

Applications Demonstration presents three practical applications of geomatics technologies in recording georeferenced EPPO standard assessments:

- **Continuous and Discrete Variables**: Investigation of minimum dataset requirements for object detection in plant counting ("On the

Minimum Dataset Requirements for Fine-Tuning an Object Detector for Arable Crop Plant Counting: A Case Study on Maize Seedlings" published in MDPI Remote Sensing, DOI: 10.3390/rs17132190)

- **Ordinal Variables**: Evaluation of machine learning approaches for phytotoxicity scoring automation ("Supporting Screening of New Plant Protection Products through a Multispectral Photogrammetric Approach Integrated with AI" published in MDPI Agronomy, DOI: 10.3390/agronomy14020306)
- **Binary and Nominal Variables**: Application of anomaly detection techniques for plant deseases unsupervised classification (initial version of scientific paper still to be submitted).

Conclusions summarise the main findings and discuss practical considerations for integrating geomatics technologies into PPP evaluation protocols.

1.4 Literature Review and Research Gap

1.4.1 Current Limitations in Agricultural Statistical Design

Traditional statistical analysis of PPP trials is based on Fisher's principles of experimental design [1–3], which emphasise randomisation, replication and blocking to ensure the validity of the results. Although this approach is well established, it has significant limitations in agricultural field studies, where environmental heterogeneity is often unknown or difficult to assess

when setting up the trial because it relies mostly on human judgement [4, 5].

The experimental design components that most heavily rely on human judgment are block disposal and control arrangement [6, 7]. Block disposal should minimize environmental variability within blocks while maximizing it between blocks [8, 9], while control arrangement ensures untreated controls are not influenced by adjacent treated plots [10].

Problems arise when environmental variability that was not previously assessed invalidates the parametric statistical analysis because this results in residual heteroscedasticity [11, 12]. Shifting to non-parametric tests often results in decreased statistical power [13, 14], so it is not a total solution.

1.4.2 Geostatistical Approaches in Agricultural Research

Geostatistical methods offer the potential to capture environmental variability mathematically rather than requiring human judgement [15, 16]. This approach should result in more robust and reliable statistical analysis, free from unexpected heterogeneity within experimental blocks [17, 18].

Several studies have demonstrated the effectiveness of spatial variability estimation through geostatistical approaches [19–23]. However, these approaches require georeferenced observations, which are often impractical for traditional visual assessments. Although some geomatics-related digital technologies can provide georeferenced observations, few studies have been conducted to demonstrate their use in experimental trials that

meet EPPO standards.

1.4.3 Digital Technologies in Agricultural Assessment

EPPO has recognised the potential of digital technologies by publishing PP 1/333(1) [24], which sets out guidelines for incorporating digital tools into PPP trials. This standard stipulates that digital data must adhere to the same quality benchmarks as manual assessments, with specific validation criteria depending on the type of variable. This allows data to be recorded in a georeferenced digital format, enabling the feasible use of geostatistical methods for PPP trial analysis.

1.4.4 Research Gap and Innovation

Despite the clear benefits of integrating geostatistics into PPP efficacy trials, the requirement for spatial coordinates alongside observations remains a significant barrier to widespread adoption. While modern technologies, toghether with geomatics techniques can provide spatial coordinates along with observations, systematic evaluation of their implementation requirements within the EPPO framework has not been conducted.

This research addresses this gap by investigating the minimum requirements needed to effectively produce georeferenced observations during efficacy trials, enabling the use of geostatistical methods to separate environmental variability from treatment effects in statistical analyses of PPP trials conducted under EPPO standards.

Theoretical Background and Methodology

2.1 Regulatory Framework

2.1.1 PPPs and EPPO Standards

PPPs are designed primarily to maintain crop health and prevent destruction by diseases and infestations. While the term "pesticides" is broader and also includes biocidal products used to control harmful organisms and disease carriers not related to plant protection, PPPs are specifically used to control harmful organisms affecting cultivated plants (such as insects, mites, fungi, bacteria, rodents, etc.), eliminate weeds, and regulate plant physiological processes. Fertilizers, which serve for plant nutrition and soil fertility improvement, are excluded from PPPs.

PPPs contain at least one active substance, which can be either chemical compounds or microorganisms, including viruses, that enable the prod-

uct to perform its intended function. These active substances undergo rigorous risk assessment processes, with EFSA (European Food Safety Authority) playing a central role in conducting peer reviews at the EU level to determine if these products, when used correctly, might produce harmful effects on human or animal health, either directly or indirectly through drinking water, food, or feed.

The main categories of PPPs can be distinguished based on the type of organism they target or the function they perform, including:

- Fungicides
- Insecticides
- Acaricides
- Nematicides
- Herbicides
- Plant growth regulators

The parameters identified through the risk assessment are compared with the values established by directive 97/57/EC [27], which indicates the acceptability limits for decision-making on the inclusion of active substances in the EU list (Annex I of directive 91/414/EEC [28]).

The Introduction of a product in the EU market is not only subject to audits on active substances and their safety for humans and environment but also to the evaluation of the product's efficacy and safety for the crop. World Trade Organization Sanitary and Phytosanitary Measures Agreement [29] recognizes the International Plant Protection Convention (IPPC)

as the only international institution in charge of emitting standards for plant health [30]. IPPC is organized in regions. European Union (EU) countries refer to the European and Mediterranean Plant Protection Organization (EPPO). EPPO Standards are divided into Standards on Phytosanitary Measures and Standards on PPPs. PPPs standards describe the efficacy evaluation of PPPs (PP 1) and good plant protection practices. EU Good Experimental Practices (GEP) units provide Biological Assessment Dossier (BAD) efficacy trials. GEP units are expected to follow EPPO PP 1 to assess PPPs selectivity detecting phytotoxicity effects, and efficacy in the complaint of Regulation (EC) No 1107/2009 of the European Parliament and Council [31].

2.2 EPPO Standards

Generics on efficacy assessments are reported in PP 1/181(5) [32], which describes herbicide, fungicide, bactericide, and insecticide efficacy on the target evaluation. PP 1/135(4) [33] describes the selectivity assessment procedures, in other words: the standard phytotoxicity assessments of PPPs. The PP 1/152 [34] standard describes the general principles for the efficacy and selectivity evaluation of PPPs, in describing the standard experimental design. Aside from the objectives of the study and the description of treatments, the PP 1/152 outlined that a comprehensive experimental design should include a description of:

- **Type of Design**
- **Sampling Method and Measures Units**

- **Statistical Analysis Plan**

2.2.1 Experimental Design

EPPO "envisage trials in which the experimental treatments are the test product(s), reference product(s) and untreated control, arranged in a suitable statistical design" [34]. The experimental design should be randomized, with replications and blocks, and should include a sufficient number of plots to ensure the statistical power of the analysis. The number of replications and blocks should be determined based on the expected variability of the data and the desired level of statistical significance in respect to control and reference treatments. The randomization of treatments within blocks should be carried out using a suitable randomization procedure to ensure that the treatments are assigned to plots in a completely random manner. The key randomization used in PPP evaluations include:

- **Completely Randomized Design (CRD)**: Treatments randomly assigned to experimental units; statistically powerful but only suitable for homogeneous trial areas where environmental variation is minimal.
- **Randomized Complete Block Design (RCBD)**: Groups plots into homogeneous blocks with each treatment appearing once per block; controls for environmental heterogeneity across the experimental area.
- **Split-Plot Design**: Used when one factor (e.g., cultivation equipment) cannot be fully randomized; creates hierarchy with whole

plots and subplots; particularly useful when plot size or equipment constraints exist.

- **Systematic designs**: Non-randomized arrangements rarely suitable for efficacy evaluations; may only be appropriate in special cases like varietal trials on herbicide selectivity.

When designing PPP trials, the arrangement of untreated controls is critical for proper efficacy assessment. According to EPPO standards, the main purpose of untreated controls is to demonstrate adequate pest infestation, without which efficacy cannot be meaningfully evaluated. Four distinct arrangements for untreated controls exist:

- **Included controls**: The most common approach, where control plots have the same shape and size as treatment plots and are fully randomized within the experimental design. This arrangement is essential when controls will be used in statistical comparisons.
- **Imbricated controls**: Control plots are arranged systematically within the trial (between blocks or between treated plots), potentially with different dimensions than treatment plots. These observations are typically not included in statistical analyses but ensure more homogeneous distribution of untreated area effects.
- **Excluded controls**: Control plots are established outside the main trial area but in similar environmental conditions. While replication is not essential, it may be beneficial in heterogeneous environments. These observations are generally excluded from statistical analyses.

- **Adjacent controls**: Each plot is divided into two subplots, with one randomly selected to remain untreated. This approach is particularly valuable in highly heterogeneous environments but requires specialized split-plot statistical analysis.

The selection of control arrangement depends on several factors: whether the control will be included in statistical tests (requiring included controls), the degree of environmental heterogeneity (adjacent controls are preferred for high heterogeneity), and the potential for control plots to interfere with adjacent treatment plots (suggesting excluded controls when interference is likely). The trials type design is critical for the success of the study, as it ensures that the results are reliable, reproducible, and statistically valid.

2.2.2 Sampling Method and Measures Units

While defining the experimental units through the randomisation design choice, The sampling method and the units of measurement must also be defined. Target and crop-specific standards point out "mode of assessment recording and measurements" fixing evaluation metrics in two ways: countable (discrete values) and measurable (continuous values) effects which must be expressed in absolute values, in other cases, frequency (incidence) and degree (severity) should be estimated and reported as affected percentage of the individual (ex. plant or plot) or as proportion within treatment and control expressed in percentage. As specified by PP 1/152 [34], classification by ranking (ordinal) and scoring (ordinal or nominal) is also contemplated. In the case of estimation, rather than count or measure, PP 1/152 reports "The observer should be trained to make

the estimations and his observations should be calibrated against a standard". Calibration compliance with standards is ensured by GEP audits. Scoring and ranking scales examples are published on specific standards or the same PP 1/152. The lack of specific scales lets trial protocol authors define one inspired in range and intervals by the mentioned examples or other well-established ones. GEP units PP 1 assessments are produced by trained and experienced agronomists or biologists by visual inspection or laboratory analysis. The technician follows the trial protocol and related EPPO standards during assessment execution. The technician is critical for accuracy, precision, and repeatability. Sensitivity is determined by the trial protocol. It depends on expected differences and if a measure, a proportion, or a scale is used. For instance, in PP 1/93(3) [35] "Efficacy evaluation of herbicides - Weeds in cereals - Observation on the crop", phytotoxicity color modification could be measured, or estimated as proportion in respect to the untreated, or scored in EPPO scale as PP 1/135(4) reports, or a scientifically accepted score as the European Weed Research Society phytotoxicity damage score [36] and other ones. In general, data types must undergo the classification presented in Table 2.1.

Table 2.1: Different modes of observation and types of variables

Type of Variable	**Measurement**	**Ranking**	**Scoring**
Binary			X
Nominal			X
Ordinal		X	X
Discrete	X		
Continuous limited	X		
Continuous not limited	X		

2.2.3 Statistical Analysis

While PP 1/152 [34] doesn't prescribe specific analyses for all situations, it emphasizes that analysis methods should align with the experimental design and data types collected. For quantitative variables (continuous or discrete), parametric methods based on Generalized Linear Models (GLM) are recommended, including ANOVA and regression approaches. For qualitative variables (bynary, ordinal or nominal), non-parametric methods are more appropriate. Parametric analysis assumes additivity of effects, homogeneity of variance, and normally distributed errors. When these assumptions aren't met, data transformations or alternative approaches become necessary.

Statistical tests, particularly F-tests of orthogonal contrasts, should focus on biologically relevant comparisons specified during the design stage: untreated control versus treatments (establishing trial validity), reference products versus control (demonstrating coherence), test products versus reference (evaluating efficacy), and comparisons among test products (identifying superior treatments). For efficacy trials, EPPO suggests one-sided tests since the aim is comparing products against references or controls, with appropriate multiple comparison procedures when needed.

In pragmatic undertakings, analyses with parametric models are frequently favoured over non-parametric ones because parametrics have greater statistical power for the same number of observations [37]. This often results in experiments being designed that use continuous or discrete variables instead of binary, nominal or ordinal ones. These continuous or discrete variables are often recorded through visual estimation rather

than true measurements or counting because the latters are more time-consuming and require more equipment [38–40]. However, using visual estimates often results in fitting parametric models that violate key assumptions, such as normally distributed residuals and homoscedasticity [41–43]. This can lead to unjustified transformations or the use of non-parametric models, resulting in a loss of statistical power.

2.2.4 Geomatics Methods in EPPO Standards

Through adherence to rigorous design, variables types and analysis standards, researchers can generate reliable evidence to support PPP registration while ensuring that products demonstrate consistent efficacy across relevant agricultural conditions. Nevertheless, the traditional statistical approaches face significant limitations in agricultural field studies where environmental heterogeneity is often unknown and difficult to predict, or even when visual estimates are used to assess variables that should be measured or counted. This is where geostatistical methods can provide a significant advantage by enabling post-hoc estimation of spatial environmental variability. The fundamental advantage of geomatics over traditional statistical approaches lies in its ability to agnostically capture spatial variability in a continuous dimension rather than predefined block and replication factors that rely on subjective experience. Traditional statistical analysis for PPP trials still relies on Fisher's principles of experimental design [1–3], which emphasize the importance of randomization, replication, and blocking to ensure the validity of results. Even if this approach for the experimental design is well-established, it is not without weaknesses: it relies on the agronomist-experimentalist knowledge and experience of

the field where the trial is performed that can be limited and biased as any human observation [4, 5]. The experimental design part that mostly relies on human choice is the block/replication disposal and the experimental units arrangement [6, 7]. The block and replication disposal should guarantee that the environmental variability is minimized within the block and maximized between blocks [8, 9], while the experimental units arrangement ensures that the treatments and untreated control are not influenced by adjacent treated plots [10]. Problems arise when environmental variability effects, unobserved during set-up, make the parametric statistical analysis invalid due to heteroscedasticity of the residuals [11, 12]. Often in such cases, shifting to non-parametric tests could mean a decrease of power [13, 14] and often is not the final solution for biased trials. If the experimental design and the statistical model could be able to catch the environmental variability "a posteriori" instead of guessing its distribution "a priori" [15, 16], the statistical analysis should be more robust, reliable and free from unexpected heterogeneity within the blocks [17, 18, 44].

EPPO recognizes the application of geostatistical methods in agricultural experimentation recommending GLMs as the underlying model structure for ANOVA. Geostatistical methods can be incorporated into Generalized Linear Mixed Models (GLMMs) to account for spatial variability arising from replication structures or pseudoreplication [65, 66], thereby improving the accuracy of inferences and reducing Type I errors [11, 44–47]. However, in order to apply geostatistical methods, experimental observations must be georeferenced, i.e. associated with spatial coordinates. This can increase field data acquisition time and the need for technological instrumentation (e.g. GPS, remote sensors or drones), affecting the logistical costs of experimentation [48–50]. Nevertheless, the evolution

of automated digital surveying technologies, such as multispectral sensors mounted on UAVs, agricultural IoT devices and georeferenced mobile apps, is rapidly reducing the cost and complexity of georeferencing in experimental agriculture [51, 52]. Additionally, using real georeferenced measurements instead of subjective visual estimates reduces the risk of violating the assumptions of parametric statistical models (e.g. normality of residuals or homogeneity of variance) [41–43]. Consequently, geostatistics can be a viable alternative to classical block-based statistics, offering greater robustness and statistical power, particularly in contexts of high spatial heterogeneity.

2.3 Geomatics

Geomatics is an integrated approach to the measurement, analysis, management, and display of geographically referenced information [53]. This research focus on two main aspects of geomatics: the acquisition of spatially referenced data and the application of geostatistical methods to analyze this data. Geomatics encompasses a range of technologies and techniques that enable precise spatial data acquisition, including georeferenced sensors, photogrammetry, spectral imaging, and machine learning. In agricultural research, georeferenced sensors integrated into machinery such as harvesters and sprayers are widely employed to collect spatially explicit data on crop yield, soil characteristics, and other agronomic parameters. These technologies are complemented by photogrammetry, which enables geometric reconstruction and the quantification of structural features from images; spectral imaging, which captures

reflectance across multiple wavelengths to infer biological and physical attributes; and machine learning algorithms, which support automated measurement, score and rank. These techniques collectively provide the capability to generate spatially-referenced datasets essential for implementing geostatistical methods in PPP trials.

2.3.1 Geostatistics

Variograms are fundamental tools in geostatistics that quantify the spatial dependence of a random field [54, 55]. They characterize how data similarity changes with distance, making possible to include the indipendent from the treatments spatial variation in a parametric model. The empirical (sample) variogram $\hat{\gamma}(h)$ is calculated by:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} [z(s_i) - z(s_j)]^2$$

where $z(s_i)$ is the observed value at location s_i, $N(h)$ is the set of all point pairs separated by distance h, and $|N(h)|$ is the number of such pairs [56, 57]. Variogram estimation typically follows a two-step process: first calculating the empirical variogram from observed data points, then fitting a theoretical model to this empirical structure [15, 16]. The empirical variogram is influenced by several factors:

- **Lag distance**: The distance interval h at which pairs of points are compared
- **Binning**: How distance classes are discretized when calculating average semivariance

- **Directional parameters**: Whether to consider anisotropy (directional dependence)
- **Maximum distance**: The upper limit of separation distance to include

Variogram parameters that require configuration, often without direct optimization so reling on statistician-agronomic knowledge [58], include:

- **Nugget** (c_0): The y-intercept of the variogram, representing measurement error and microscale variation
- **Sill** (c): The plateau value reached by the variogram, equal to the variance of the random field
- **Range** (a): The distance beyond which observations become spatially independent
- **Anisotropy ratio**: The ratio between the maximum and minimum ranges in different directions
- **Anisotropy angle**: The direction of maximum spatial continuity

Theoretical variogram models must be selected to fit the empirical variogram [54, 55]. Common models include:

- **Spherical model**: Exhibits a progressive decrease of spatial dependence until reaching the range
- **Exponential model**: Approaches the sill asymptotically, with practical range typically defined at 95% of the sill

- **Gaussian model**: Shows a parabolic behavior near the origin, indicating high continuity
- **Matérn model**: Offers flexibility through an additional smoothness parameter
- **Power model**: Used for non-stationary processes without a finite variance
- **Nugget effect model**: Represents microscale variation or measurement error

Proper variogram modeling for excluding spatial variation in a parametric model must be fitted on control samples [44, 59, 60] to avoid including treatment effects in the spatial terms of the parametric model. The control samples must be homogeneously distributed in the space and also in time if the aim is to get spatiotemporal variation. In order to ensure the right control sampling, it is necessary to implement a combined approach using both imbricated or adjacent controls along with included controls in the following manner: imbricated or adjacent control observations to fit the variogram model while the included control to test the error of the variogram predictions in a cross-validation fashon. Testing on included controls ensures the variogram model error estimates aren't biased by the spatial arrangement of control observations. Nevertheless, it is possible to not include control and test variogram model errors through other cross-validation technics such as Leave-One-Out or K-Folds on the imbricated or adjacent controls [16, 61]. If the only imbricated control is present, area size of control is even more important, because having it smaller than treatment plot size could lead to a wrong estimation of the

variogram model. In practical terms, in an imbricated control arrangement, a buffer area with width equal to the treatment plot width must be placed around each plot, resulting equivalent to an adjacent control arrangement. The spatial varibility estimation through variogram modeling was already proved to be effective in many studies [19–23], but it has the drawback of requiring a large control area.

If incorporating such large control areas is not feasible and only included controls are available, one can rely on Spatial Analysis of field Trials with Splines (SpATS) [62–64]. SpATS is a statistical model that enables correction of spatial heterogeneity of the data by using splines to model the spatial trend of the response variable. The SpATS model is based on the assumption that the response variable can be described as a function of the treatment effects and a smooth spatial trend. The model can be expressed as:

$$y_{ij} = \mu + \tau_i + f(u_j, v_j) + \varepsilon_{ij}$$

where:

- y_{ij} is the response variable observed at the j-th location for the i-th treatment
- μ is the overall mean
- τ_i represents the fixed effect of the i-th treatment
- $f(u_j, v_j)$ is the smooth spatial trend modeled using tensor-product P-splines, where u_j and v_j are the spatial coordinates of the j-th location

- ε_{ij} is the random error term, typically assumed to be normally distributed with zero mean and constant variance

The spatial component $f(u_j, v_j)$ can be further decomposed into additive and interaction effects:

$$f(u_j, v_j) = f_1(u_j) + f_2(v_j) + f_{12}(u_j, v_j)$$

where $f_1(u_j)$ and $f_2(v_j)$ are the main effects for rows and columns, and $f_{12}(u_j, v_j)$ represents the smooth interaction surface that accounts for localized spatial patterns. The smoothing parameters controlling the flexibility of these components are estimated using restricted maximum likelihood (REML). According to Rodriguez-Alvarez et al. (2018) [62], SpATS has been effectively applied to field trials with varying dimensions, but there are practical considerations for minimum requirements: a minimum grid of 5x5 (25 plots) is often considered a practical lower limit, but complex spatial variations might require a bigger grid. SpATS has been successfully applied to breeding trials with experimental designs exceeding hundreds of rows and columns, whereas the variogram approach can be effective even with a single treatment plot, provided sufficient control area is available. Thus, as a rule of thumb, for trials with sufficient space to implement imbricated or adjacent checks with included control, variogram is advised, while for trials with many individuals (at least more than 25 in a regular grid) and constrained space, SpATS estimations is often better. To evaluate how effectively the fitted model describes spatial variability of the data for SpATS, residuals of included control repetitions (compared to their mean) should be smaller than the combined model random and unexplained variation. For both approaches, finding the optimal model

across parameters that cannot be directly solved requires, an iterative approach testing various settings can be deployed.

All the described geostatistical techniques can be seamlessly integrated with GLMMs to provide a comprehensive analysis of spatial-temporal data [62, 67], enhancing the accuracy and reliability of treatment effect estimates in orthogonal tests.

2.3.2 Geomatics Thecnics and Digital Technologies

As already mentioned, the EPPO standards require that the observations are made by trained experimentalists. To regulate the use of technologies that can measure, rank or score instead of experimentalists, the EPPO published a new standard (PP 1/333(1)[24]) which addresses the use of digital technologies in PPP efficacy and selectivity trials. This standard provides guidelines for incorporating digital tools into trial protocols, where digital tools are intended as a combination of hardwares and softwares delivering data in a semi-automatic or automatic fashon. The digital data must respect the same quality standards of the manual ones, and the digital tools must be validated before the trial execution. Validation of digital tools should be performed by comparing the results of digital and manual assessments, demonstrating that the digital tools provide reliable and consistent results compared to manual assessments golden sample. The benchmarks for the validation depends on the type of variable. For each type of variable, the congruence between digital and manual should be evaluated with a different metric:

- **Continuous and Discrete**: Coefficient of determination (R^2) higher

than 0.85 (when fitting a 1:1 relationship model (y = x) between digital and manual measurements).

- **Ordinal and Nominal**: Cohen's kappa Coefficient (κ) higher than 0.7.
- **Binary**: Accuracy higher than 0.85

The variable type also influence the kind of digital tool to use. The hardware of a digital tool is always a sensor to collect the raw data and a processor to convert the raw data in a digital format. For what concerns the software of a digital tool, it is worth to mention that the core of it is always a model that convert the digital format into the assessment observation in the variable units. Quantitative variables are produced by regression models, while qualitative (categorical) variables are produced by classification models. Quantitative variables: continuous (limited or not) and discrete can be summarized as metric measurments and counts respectively.

The capability of digital tools to measure and georeference harvest-related parameters such as yield weight, moisture content, and hectolitric weight is well established [50, 51, 86]. However, parameters that are more complex to assess digitally, including plant counts, phytotoxicity symptoms, and disease incidence or severity, have garnered significant research interest in recent decades [86, 97, 98, 101]. Despite technological advancements, a definitive consensus has yet to emerge regarding the extent to which these tools can fully substitute for human observers [97, 99, 100]. Nevertheless, their adoption undoubtedly enhances field assessments by incorporating georeferenced data, so enabling geostatistics for efficacy trials. This study aims to evaluate the potential of geomatics-based tech-

nologies for collecting observational data in agricultural trials conducted under EPPO standards. Specifically, it investigates the use of photogrammetry, spectral imaging, and machine learning as digital tools for capturing trial observations across all variable types recognized by EPPO guidelines.

Photogrammetry

Photogrammetry is a technique used to obtain reliable information about physical objects and the environment through the process of recording, measuring, and interpreting photographic images. It is widely used in various fields such as topographic mapping, architecture, engineering, manufacturing, quality control, and geology. The fundamental principle of photogrammetry is based on the geometry of image formation and the mathematical relationships between the images and the objects being photographed [68, 69]. The basic principle of photogrammetry involves capturing multiple photographs of an object or scene from different perspectives. By analyzing these images, it is possible to reconstruct the three-dimensional (3D) coordinates of points on the object's surface. The key steps in photogrammetry include image acquisition, image orientation, and 3D reconstruction [68, 70].

Images are typically captured using cameras mounted on various platforms such as tripods, drones, or aircraft. The quality and resolution of the images are crucial for accurate photogrammetric analysis. The images should have sufficient overlap (usually 60-80%) to ensure that common points are visible in multiple images [71, 72].

Image orientation involves determining the position and orientation of the

camera at the time each photograph was taken. This process is divided into two main steps: interior orientation and exterior orientation [73, 74].

- **Interior Orientation:** This step involves determining the internal geometry of the camera, including the focal length, principal point, and lens distortion parameters. These parameters are typically obtained through a camera calibration process.
- **Exterior Orientation:** This step involves determining the position (X, Y, Z coordinates) and orientation (roll, pitch, yaw angles) of the camera in a global coordinate system. While tie points from overlapping images establish relative orientation between images, absolute exterior orientation requires Ground Control Points (GCPs) with known coordinates or other external reference information (such as GNSS/IMU data) to establish the relationship between image coordinates and real-world coordinates.

Once the images are oriented, the 3D coordinates of points on the object's surface can be reconstructed using triangulation. Triangulation is a mathematical process that involves intersecting lines of sight from multiple images to determine the precise location of a point in 3D space [68, 75].

Mathematically, the process can be described using the collinearity equations, which relate the image coordinates (x, y) of a point to its object coordinates (X, Y, Z) through the camera parameters [69, 76]:

$$x = x_0 - \frac{f \cdot (r_{11}(X - X_0) + r_{12}(Y - Y_0) + r_{13}(Z - Z_0))}{r_{31}(X - X_0) + r_{32}(Y - Y_0) + r_{33}(Z - Z_0)}$$
$$y = y_0 - \frac{f \cdot (r_{21}(X - X_0) + r_{22}(Y - Y_0) + r_{23}(Z - Z_0))}{r_{31}(X - X_0) + r_{32}(Y - Y_0) + r_{33}(Z - Z_0)}$$

where:

- (x_0, y_0) are the coordinates of the principal point in the image.
- f is the focal length of the camera.
- (X_0, Y_0, Z_0) are the coordinates of the camera position.
- r_{ij} are the elements of the rotation matrix that describes the orientation of the camera.

By solving these equations for multiple images, the 3D coordinates of the object points can be accurately determined. Recognized object points can be more or less sparse depending on the approach to their recognition and pairing. Historically, homologous points within images was performed manually. Today many algorithms for this task are available. They can be divided between point-based and area-based algorithms. Point-based algorithms identify and match distinct points in the images, such as corners or edges. These algorithms rely on feature descriptors (e.g., SIFT, SURF, ORB) [77–79] to extract and match key points across images. Area-based algorithms, on the other hand, use the entire image region to find correspondences. They typically involve template matching or correlation techniques to identify similar regions in different images. The choice of algorithm depends on the specific application and the characteristics of the images being processed. The 3D reconstruction process can also be enhanced using additional techniques such as structure from motion (SfM) and multi-view stereo (MVS) [80, 81]. SfM is a technique that estimates the camera motion and 3D structure of a scene simultaneously from a set of images. It involves detecting and matching feature

points across multiple images, and then using these correspondences to estimate the camera parameters and the 3D coordinates of the points. MVS, on the other hand, focuses on dense reconstruction by estimating the depth information for each pixel in the images. It uses the camera parameters and the matched feature points to create a dense point cloud or a 3D mesh of the scene. The resulting 3D model can be visualized and analyzed using specialized software, allowing for measurements of distances, areas, and volumes. Photogrammetry can also be used to create orthophotos, which are geometrically corrected images that can be used for mapping and analysis [82, 83]. Orthophotos are generated by removing the effects of perspective distortion and terrain relief from the original images, resulting in a scale-accurate representation of the area.

Spectral Imaging

Spectral imaging is a technique that captures images at multiple wavelengths of the electromagnetic spectrum, providing valuable information about the spectral characteristics of objects and materials [84, 85]. Multispectral images are typically acquired using specialized cameras or sensors that can capture light in different spectral bands, ranging from ultraviolet wavelengths (UV, 200-380 nm) to near infrared wavelengths (NIR 750-1400 nm) or short-wavelength infrared (SWIR 1.4-3 μm) [85, 86]. Some study tested also the feasibility to use bands in the termal infrared (TIR, 3-15 μm) range [87]. Each spectral band corresponds to a specific range of wavelengths, allowing for the analysis of the spectral signature of objects in the scene. The spectral signature is the unique pattern of reflectance or absorption of light at different wavelengths for a specific ma-

terial or object. By analyzing the spectral signatures of different materials, it is possible to identify and classify them based on their spectral characteristics. This is particularly useful in applications such as vegetation analysis, where different plant species exhibit distinct spectral signatures due to variations in leaf pigments, moisture content, and other physiological factors [88]. Multispectral imaging can be performed using various platforms, including drones, satellites, and ground-based systems. The choice of platform depends on the specific application, the spatial resolution required, and the area of interest. The images captured by multispectral sensors are typically processed using specialized software that applies various algorithms to extract relevant information from the spectral data. This processing may include radiometric correction, geometric correction, and atmospheric correction to ensure accurate and reliable results. The resulting multispectral images can be analyzed to derive various indices and metrics that provide insights into the health and condition of vegetation. One of the most commonly used indices in agriculture is the Normalized Difference Vegetation Index (NDVI), which is calculated using the red and NIR bands of the multispectral image, usually with 650-680 nm and 855-875 nm bands respectivelly [89, 90]. NDVI is a measure of vegetation greenness and is widely used to assess plant health, monitor crop growth, and detect stress conditions. The NDVI is calculated using the following formula:

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

where NIR and Red are the reflectance values in the near-infrared and red bands, respectively. NDVI values range from -1 to +1, with higher val-

ues indicating greater vegetation density and health. NDVI is particularly useful for monitoring crop growth, assessing drought conditions, and detecting plant diseases. Other indices derived from multispectral images include the Enhanced Vegetation Index (EVI) and Soil-Adjusted Vegetation Index (SAVI) each providing specific information about vegetation health and condition [91, 92]. These indices can be used to monitor crop performance, assess nutrient status, and evaluate the impact of environmental factors on plant growth [86, 93]

Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) that focuses on the development of algorithms and models capable of learning from and making predictions or decisions based on data [94]. Unlike traditional programming, where explicit instructions dictate the output for given inputs, ML models identify patterns and relationships within data to generate predictive outcomes [95]. These techniques are particularly valuable when dealing with large, complex, or high-dimensional datasets, where manual analysis would be impractical or inefficient.

ML has gained substantial importance in various scientific fields, including agriculture [96] and plant protection [97]. Within the context of PPP efficacy evaluation, ML offers new opportunities to enhance data processing, interpretation, and decision-making by leveraging vast amounts of observational data collected during field trials [98]. Integrating ML approaches into the framework of PP1/333 can significantly increase the robustness and accuracy of the analysis, allowing for more data-driven and automated assessments [99–101].

The primary objective of employing ML techniques in PPP trials is to improve accuracy, precision, and reproducibility while reducing manual intervention and subjective bias [97]. Modern ML methods can analyze complex interactions between variables and predict treatment outcomes under various conditions, thereby facilitating more efficient and accurate efficacy assessments.

There are several fundamental approaches in machine learning, each suited to different types of tasks and data structures:

- **Supervised Learning**: Models are trained on labeled datasets where the input-output relationship is known. Techniques include regression, classification, and ensemble methods such as Random Forests and Gradient Boosting.
- **Unsupervised Learning**: Models identify patterns or groupings within data without labeled responses. Clustering (e.g., K-means, hierarchical clustering) and dimensionality reduction (e.g., PCA, t-SNE) are common techniques.
- **Weakly-supervised Learning**: Combines a small amount of labeled data with a large amount of unlabeled data to improve learning accuracy.
- **Self-supervised Learning**: A machine learning approach where the model generates its own labels from the input data, creating supervised-like learning tasks without external human annotations. The model learns by solving tasks designed within the data itself, such as reconstructing partially obscured images. This technique allows models to learn rich, generalizable representations from large

unlabeled datasets, enabling transfer learning and reducing the dependency on expensive manual labeling.

ML models can also be integrated with statistical techniques, providing hybrid approaches that combine inferential statistics with predictive modeling [95]. For example, generalized linear models (GLMs) can be enhanced with ML techniques to improve their accuracy and adaptability [102]. Deep Learning (DL) is a subfield of ML that studies a particular class of models named Deep Neural Networks [103], the most active ML study area since ten years. Computer vision (CV) is another subfield of ML that focuses on enabling machines to interpret and analyze visual information. CV is the more and more treated lonely with DL instead of other ML approaches. In the context of PPP efficacy evaluation, computer vision methods are increasingly used for automated observation and measurement, particularly when integrated with digital imaging and photogrammetry [104]. The use of computer vision within PP1/333 trials significantly enhances data acquisition by enabling digital sensing and precise measurement of crop conditions [99, 100]. Techniques such as image segmentation, object detection, and texture analysis can automatically identify plant stress, disease symptoms, and pest damage [26]. In the context of experimental tests, object detection and segmentation serve to localize in space and time the observations used during statistical tests on models such as GLMs. This allows the exclusion of spatiotemporal variability as explained in the next chapter 'Geostatistics'. Moreover, combining computer vision with geostatistical methods allows for the spatial mapping of efficacy across field plots, generating comprehensive visual assessments that support statistical evaluations [105].

Having representative big datasets is a significant challenge in DL. Large-scale, high-quality datasets are crucial for training robust machine learning models, but acquiring such datasets is often prohibitively expensive and time-consuming. Many domains, particularly specialized fields like PPPs and phytopathometry research, struggle to compile sufficiently large and diverse training datasets [106]. The data collection process in Supervised Learning involves manual annotation, which introduces human bias and can be extremely labor-intensive. Moreover, ensuring dataset representativeness is complex, as minor sampling biases can lead to models that perform poorly when deployed in real-world scenarios [107]. Weakly-supervised and Self-supervised Learning came to leverage this problem giving the possibility to train models with respectivelly few or without human supervision. Weakly-supervised learning leverages pre-trained models developed through enormous computational efforts, resulting in foundation models with remarkable generalization capabilities [108]. These models can effectively perform new tasks with minimal fine-tuning, a phenomenon known as "few-shot learning" or "in-context learning" [109]. The remarkable ability of these models to adapt to new tasks with very few examples represents a paradigm shift in machine learning, where the pre-training phase becomes crucial in developing adaptable and versatile AI systems [110]. Self-supervised learning, while promising to revolutionize machine learning by eliminating the need for manual labeling, presents its own set of challenges [111]. The computational resources required for training large self-supervised models are substantial, often exceeding the capabilities of smaller research labs or specialized studies [112]. This computational intensity creates a significant barrier to entry, particularly for domain-specific research like PPP efficacy evaluation, where

the computational and expertise requirements may outstrip the available resources of a typical research group [113]. Despite its transformative potential, self-supervised learning remains a cutting-edge approach that requires significant computational infrastructure and interdisciplinary expertise to implement effectively.

A way to use these advanced learning approaches is to leverage pre-trained models as feature extractors through unsupervised inference techniques [114]. Researchers can exploit the rich representations learned by foundation models, applying them as powerful feature extraction mechanisms across various downstream tasks [115]. Alternatively, these pre-trained models can serve as robust backbones, with researchers fine-tuning only the final classification or prediction layers to adapt the model to specific domain requirements [111]. This transfer learning approach allows for efficient model adaptation, reducing the need for extensive domain-specific data collection and annotation [116]. In the context of specialized fields like phytosanitary research, such techniques enable more efficient model development by leveraging the generalization capabilities of large-scale pre-trained models, effectively bridging the gap between computational limitations and domain-specific research needs [117]. The transfer of knowledge from foundation models to domain-specific applications represents a significant advancement in machine learning methodologies. By extracting and repurposing learned representations, researchers can develop more sophisticated and adaptable models with minimal additional training resources [118]. This approach not only mitigates the challenges of data scarcity but also provides a more computationally efficient pathway to developing advanced predictive models in specialized research domains [119].

2.4 The Literature Gap and the Thesis Aims

Geostatistics has been shown to estimate environmental variation more effectively than randomization because it relies on mathematical modeling rather than field technician experience [44]. Despite the clear benefits of integrating geostatistics into phytosanitary product efficacy trials, the requirement for spatial coordinates alongside observations remains a barrier to widespread adoption [15, 16]. By leveraging geomatics techniques such as georeferenced sensors, photogrammetry, spectral imaging, and ML, researchers can improve data collection and analysis, as their synergy can provide spatial coordinates along with the observations. Through a series of case studies addressing each variable type described in PP 1/152 [34] and PP 1/333(1) [24], this work explores the opportunities and limitations of deploying geomatic techniques to improve PPP effect estimation. First, the research describes the use of photogrammetry and ML to localize observations along with plant counting, a typical continuous variable assessment. After demonstrating the viability of obtaining samples and/or subsamples spatial coordinates in a geographical coordinates reference system, the study establishes the practicality of using a local coordinates reference system to deploy geostatistics in protected cultures such as in a greenhouse. Thus, in the second case study, the research evaluates the effectiveness of reproducing an ordinal scale assessment like the phytotoxicity scoring with ML and discusses whether digitalization can replace ordinal scale values by converting them to continuous scale values for implementing parametric tests instead of not parametric ones. Finally the work tests the applicability of using already trained ML models on georeferenced samples or subsamples. This approach will be used

to obtain binary or nominal variable assessments, classifying between healthy or unhealthy plants, or even between different diseases. These classifications will be accomplished without specific task-oriented training, but rather through the implementation of anomaly detection and unsupervised learning techniques. In summary, this collection of case studies aims to determine the minimum requirements needed to effectively implement geomatic techniques during an efficacy trial. These techniques will enable the use of geostatistical methods to separate environmental variability from treatment effects in statistical analyses of PPP trials conducted under the EPPO standard framework.

Bibliography

[1] R. A. Fisher. Statistical Methods for Research Workers. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 66–70. Springer, New York, NY, 1992.

[2] R. Mead, S. G. Gilmour, and A. Mead. *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. Cambridge University Press, September 2012.

[3] Michael D. Casler. Fundamentals of Experimental Design: Guidelines for Designing Successful Experiments. *Agronomy Journal*, 107(2):692–705, 2015.

[4] Giuditta Parolini. In pursuit of a science of agriculture: The role of statistics in field experiments. *History and Philosophy of the Life Sciences*, 37(3):261–281, September 2015.

[5] Dominic Berry. The resisted rise of randomisation in experimental design: British agricultural science, c.1910-1930. *History and Philosophy of the Life Sciences*, 37(3):242–260, September 2015.

[6] K. D. Tocher. The Design and Analysis of Block Experiments.

Journal of the Royal Statistical Society: Series B (Methodological), 14(1):45–91, 1952.

[7] Emlyn Williams and Hans-Peter Piepho. Optimality and Contrasts in Block Designs with Unequal Treatment Replication. *Australian & New Zealand Journal of Statistics*, 57(2):203–209, 2015.

[8] H. M. van Es and C. L. van Es. Spatial Nature of Randomization and Its Effect on the Outcome of Field Experiments. *Agronomy Journal*, 85(2):420–428, 1993.

[9] C. J. Brien, B. D. Harch, R. L. Correll, and R. A. Bailey. Multiphase Experiments with at Least One Later Laboratory Phase. I. Orthogonal Designs. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(3):422–450, September 2011.

[10] H. P. Piepho, J. Möhring, and E. R. Williams. Why Randomize Agricultural Experiments? *Journal of Agronomy and Crop Science*, 199(5):374–383, 2013.

[11] Oliver Schabenberger and Carol A. Gotway. *Statistical Methods for Spatial Data Analysis*. CRC Press, December 2004.

[12] A Onofri, F Gresta, and F Tei. A new method for the analysis of germination and emergence data of weed species. *Weed Research*, 50(3):187–198, 2010.

[13] Walter W. Stroup. Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. *Agronomy Journal*, 107(2):811–827, 2015.

[14] Alan Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, November 2018.

[15] M.A. Oliver, editor. *Geostatistical Applications for Precision Agriculture*. Springer Netherlands, Dordrecht, 2010.

[16] Richard Webster and Margaret A. Oliver. *Geostatistics for Environmental Scientists*. John Wiley & Sons, October 2007.

[17] Christel Richter and Bärbel Kroschewski. Geostatistical Models in Agricultural Field Experiments: Investigations Based on Uniformity Trials. *Agronomy Journal*, 104(1):91–105, 2012.

[18] María V. López and José L. Arrúe. Efficiency of an Incomplete Block Design Based on Geostatistics for Tillage Experiments. *Soil Science Society of America Journal*, 59(4):1104–1111, 1995.

[19] David S. Bullock, Maria Boerngen, Haiying Tao, Bruce Maxwell, Joe D. Luck, Luciano Shiratsuchi, Laila Puntel, and Nicolas F. Martin. The Data-Intensive Farm Management Project: Changing Agronomic Research Through On-Farm Precision Experimentation. *Agronomy Journal*, 111(6):2736–2746, 2019.

[20] A. Castrignanò, R. Quarto, A. Venezia, and G. Buttafuoco. A geostatistical approach for modelling and combining spatial data with different support. *Advances in Animal Biosciences*, 8(2):594–599, January 2017.

[21] Huidong Jin, K. Shuvo Bakar, Brent L. Henderson, Robert G. V. Bramley, and David L. Gobbett. An efficient geostatistical analysis tool for on-farm experiments targeted at localised treatment. *Biosystems Engineering*, 205:121–136, May 2021.

[22] Laila A. Puntel, Laura J. Thompson, and Taro Mieno. Leveraging digital agriculture for on-farm testing of technologies. *Frontiers in Agronomy*, 6, March 2024.

[23] R. G. Trevisan, D. S. Bullock, and N. F. Martin. Spatial variability of crop responses to agronomic inputs in on-farm precision experimentation. *Precision Agriculture*, 22(2):342–363, April 2021.

[24] PP 1/333 (1) Adoption of digital technology for data generation for the efficacy evaluation of plant protection products. *EPPO Bulletin*, n/a(n/a).

[25] Anne-Katrin Mahlein. Plant Disease Detection by Imaging Sensors – Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. *Plant Disease*, 100(2):241–251, February 2016.

[26] Andreas Kamilaris and Francesc X. Prenafeta-Boldu. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, April 2018.

[27] European Commission. Council directive 97/57/EC of 22 september 1997, uniform Principles for evaluation and authorisation of plant protection products, 1997.

[28] Council of the European Communities. Council Directive 91/414/EEC of 15 July 1991 concerning the placing of plant protection products on the market, 1991.

[29] World Trade Organization. The WTO agreement on the application

of sanitary and phytosanitary measures (SPS agreement). *World Trade Organization*, 1995.

[30] International Plant Protection Convention. International standards for phytosanitary measures (ispms), 2022.

[31] European Parliament and Council. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market, 2009.

[32] EPPO. PP 1/181(5) Conduct and reporting of efficacy evaluation trials, including good experimental practice. Technical report, European and Mediterranean Plant Protection Organization, 2021.

[33] EPPO. PP 1/135(4) phytotoxicity assessment. Technical report, European and Mediterranean Plant Protection Organization, 2014.

[34] EPPO. PP 1/152 Design and analysis of efficacy evaluation trials. Technical report, European and Mediterranean Plant Protection Organization, 2012.

[35] EPPO. PP 1/93(3) weeds in cereals. Technical report, European and Mediterranean Plant Protection Organization, 2015.

[36] H. Bleiholder, T. van den Boom, P. Langelüddeke, and R. Stauss. Einheitliche codierung der phänologischen entwicklungsstadien mono- und dikotyler pflanzen - erweiterte BBCH-skala, allgemein. *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes*, 43:265–270, 1991.

[37] Alan Agresti. Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine*, 20(17-18):2709–2722, 2001. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.738.

[38] Clive H. Bock, Sarah J. Pethybridge, Jayme G. A. Barbedo, Paul D. Esker, Anne-Katrin Mahlein, and Emerson M. Del Ponte. A phytopathometry glossary for the twenty-first century: Towards consistency and precision in intra- and inter-disciplinary dialogues. *Tropical Plant Pathology*, 47(1):14–24, February 2022.

[39] K. S. Chiang, H. I. Liu, Y. L. Chen, M. El Jarroudi, and C. H. Bock. Quantitative Ordinal Scale Estimates of Plant Disease Severity: Comparing Treatments Using a Proportional Odds Model. *Phytopathology*, 110(4):734–743, April 2020.

[40] Wanderson Bucker Moraes, Laurence V. Madden, and Pierce A. Paul. Characterizing Heterogeneity and Determining Sample Sizes for Accurately Estimating Wheat Fusarium Head Blight Index in Research Plots. *Phytopathology*, 112(2):315–334, February 2022.

[41] M. Stevenson, M. Segui-Gomez, I. Lescohier, C. Di Scala, and G. McDonald-Smith. An overview of the injury severity score and the new injury severity score. *Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention*, 7(1):10–13, March 2001.

[42] Marco Acutis, Barbara Scaglia, and Roberto Confalonieri. Perfunctory analysis of variance in agronomy, and its consequences in ex-

perimental results interpretation. *European Journal of Agronomy*, 43:129–135, November 2012.

[43] Kuo-Szu Chiang, Shih-Chia Liu, Clive H. Bock, and Tim R. Gottwald. What Interval Characteristics Make a Good Categorical Disease Assessment Scale? *Phytopathologyő*, 104(6):575–585, June 2014.

[44] Darghan C. Aquiles E., Taborda L. Darlley S., González S. Nair J., Rivera M. Carlos A., and Ospina N. Jesús E. The effect of spatial lag on modeling geomatic covariates using analysis of variance. *Applied Geomatics*, 16(3):779–788, September 2024.

[45] Johanna I. F. Slaets, Runa S. Boeddinghaus, and Hans-Peter Piepho. Linear mixed models and geostatistics for designed experiments in soil science: Two entirely different methods or two sides of the same coin? *European Journal of Soil Science*, 72(1):47–68, 2021. _eprint: https://bsssjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/ejss.12976.

[46] Russell B. Millar and Marti J. Anderson. Remedies for pseudoreplication. *Fisheries Research*, 70(2):397–407, December 2004.

[47] Hans-Peter Piepho, Christel Richter, Joachim Spilke, Karin Hartung, Arndt Kunick, and Heinrich Thöle. Statistical aspects of on-farm experimentation. *Crop and Pasture Science*, 62:721–735, 11 2011.

[48] Vyron Antoniou. Chapter 8 - On volunteered geographic information quality: a framework for sharing data quality informa-

tion. In Nikolaos Stathopoulos, Andreas Tsatsaris, and Kleomenis Kalogeropoulos, editors, *Geoinformatics for Geosciences*, Earth Observation, pages 149–160. Elsevier, January 2023.

[49] Angelos Alexopoulos, Konstantinos Koutras, Sihem Ben Ali, Stefano Puccio, Alessandro Carella, Roberta Ottaviano, and Athanasios Kalogeras. Complementary Use of Ground-Based Proximal Sensing and Airborne/Spaceborne Remote Sensing Techniques in Precision Agriculture: A Systematic Review. *Agronomy*, 13(7):1942, July 2023. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[50] Zhan Zhao, Yaoming Li, Jin Chen, and Jiaojiao Xu. Grain separation loss monitoring system in combine harvester. *Computers and Electronics in Agriculture*, 76(2):183–188, 2011.

[51] Isabel Cisternas, Ignacio Velásquez, Angélica Caro, and Alfonso Rodríguez. Systematic literature review of implementations of precision agriculture. *Computers and Electronics in Agriculture*, 176:105626, September 2020.

[52] Borgogno Mondino and Enrico Corrado. Considerazioni su costi e mercato potenziali del telerilevamento da SAPR in Italia nel settore vitivinicolo. *AIT Conference 11ř Workshop tematico di Telerilevamento*, 2017. Accepted: 2017-06-30T09:06:36Z.

[53] Mario A. Gomarasca. *Basics of Geomatics*. Springer Science & Business Media, September 2009.

[54] Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, March 2015.

[55] Pierre Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, 1997.

[56] Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, December 1963.

[57] A. G. Journel, Andre G. Journel, and Ch J. Huijbregts. *Mining Geostatistics*. Blackburn Press, 2003.

[58] Sebastian Müller, Lennart Schüler, Alraune Zech, and Falk Heße. Geostatistical modelling in Python. *Geoscientific Model Development*, 15(7):3161–3182, April 2022.

[59] R. M Lark. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105(1):49–80, January 2002.

[60] Budiman Minasny and Alex B McBratney. The efficiency of various approaches to obtaining estimates of soil hydraulic properties. *Geoderma*, 107(1):55–70, May 2002.

[61] Tomislav Hengl. *A Practical Guide to Geostatistical Mapping of Environmental Variables*. European commission. Joint research centre. Institute for environment and sustainability (Ispra, Italie), 2007.

[62] María Xosé Rodríguez-Álvarez, Martin P. Boer, Fred A. van Eeuwijk, and Paul H. C. Eilers. Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics*, 23:52–71, March 2018.

[63] María Xosé Rodríguez-Álvarez, Dae-Jin Lee, Thomas Kneib, María Durbán, and Paul Eilers. Fast smoothing parameter separation in

multidimensional generalized P-splines: The SAP algorithm. *Statistics and Computing*, 25(5):941–957, September 2015.

[64] Dae-Jin Lee, María Durbán, and Paul Eilers. Efficient two-dimensional smoothing with P<math><mi is="true">P</mi></math>-spline ANOVA mixed models and nested bases. *Computational Statistics & Data Analysis*, 61:22–37, May 2013.

[65] Edward E. Gbur, Walter W. Stroup, Kevin S. McCarter, Susan Durham, Linda J. Young, Mary Christman, Mark West, and Matthew Kramer. *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*. John Wiley & Sons, January 2020.

[66] Levi Kumle, Melissa L.-H. Võ, and Dejan Draschkow. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, 53(6):2528–2543, December 2021.

[67] Johanna I. F. Slaets, Runa S. Boeddinghaus, and Hans-Peter Piepho. Linear mixed models and geostatistics for designed experiments in soil science: Two entirely different methods or two sides of the same coin? *European Journal of Soil Science*, 72(1):47–68, 2021.

[68] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[69] Karl Kraus. *Photogrammetry: Geometry from Images and Laser Scans*. Walter de Gruyter, 2007.

[70] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer International Publishing, Cham, 2022.

[71] I. Colomina and P. Molina. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92:79–97, 2014.

[72] F. C. Nex and F. Remondino. UAV for 3D mapping applications : A review. *Applied geomatics*, 6(1):1–2015, 2014.

[73] Duane Brown. Close-Range Camera Calibration. *Photogramm. Eng.*, 37, December 2002.

[74] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 298–372, Berlin, Heidelberg, 2000. Springer.

[75] Yasutaka Furukawa and Jean Ponce. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, August 2010.

[76] Paul R. Wolf, Bon A. DeWitt, and Benjamin E. Wilkinson. *Elements of Photogrammetry with Application in GIS, Fourth Edition*. McGraw Hill Professional, October 2013.

[77] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[78] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.

[79] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571, November 2011.

[80] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds. 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314, December 2012.

[81] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528, June 2006.

[82] Edward M. Mikhail, James S. Bethel, and J. Chris McGlone. *Introduction to Modern Photogrammetry*. John Wiley & Sons, March 2001.

[83] David P. Paine and James D. Kiser. *Aerial Photography and Image Interpretation*. John Wiley & Sons, February 2012.

[84] F. D. van der Meer and S. M. de Jong. *Imaging Spectrometry : Basic Principles and Prospective Applications*. Kluwer Academic, 2001.

[85] Prasad S. Thenkabail and John G. Lyon, editors. *Hyperspectral Remote Sensing of Vegetation*. CRC Press, Boca Raton, April 2016.

[86] A.-K. Mahlein, M.T. Kuska, J. Behmann, G. Polder, and A. Walter. Hyperspectral Sensors and Imaging Technologies in Phytopathology: State of the Art. *Annual Review of Phytopathology*, 56(Volume 56, 2018):535–558, 2018.

[87] Quanxing Wan, Magdalena Smigaj, Benjamin Brede, and Lammert Kooistra. Optimizing UAV-based uncooled thermal cameras in field conditions for precision agriculture. *International Journal of applied Earth Observation and Geoinformation*, 134:104184, November 2024.

[88] Filippo Sarvia, Samuele De Petris, Alessandro Farbo, and Enrico Borgogno-Mondino. Geometric vs spectral content of Remotely Piloted Aircraft Systems images in the Precision agriculture context. *The Egyptian Journal of Remote Sensing and Space Sciences*, 27(3):524–531, September 2024.

[89] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the Great Plains with ERTS. January 1974.

[90] Compton J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, May 1979.

[91] A Huete, K Didan, T Miura, E. P Rodriguez, X Gao, and L. G Ferreira. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1):195–213, November 2002.

[92] J. Qi, A. Chehbouni, A. R. Huete, Y. H. Kerr, and S. Sorooshian. A modified soil adjusted vegetation index. *Remote Sensing of Environment*, 48(2):119–126, May 1994.

[93] Hamlyn G. Jones and Robin A. Vaughan. *Remote Sensing of Vegetation: Principles, Techniques, and Applications*. OUP Oxford, July 2010.

[94] John R. Koza, Forrest H. Bennett, David Andre, and Martin A. Keane. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In John S. Gero and Fay Sudweeks, editors, *Artificial Intelligence in Design '96*, pages 151–170. Springer Netherlands, Dordrecht, 1996.

[95] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009.

[96] Sara Oleiro Araújo, Ricardo Silva Peres, José Cochicho Ramalho, Fernando Lidon, and José Barata. Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy*, 13(12):2976, December 2023.

[97] Clive H. Bock, Jayme G. A. Barbedo, Emerson M. Del Ponte, David Bohnenkamp, and Anne-Katrin Mahlein. From visual estimates to fully automated sensor-based measurements of plant disease severity: Status and challenges for improving accuracy. *Phytopathology Research*, 2(1):9, April 2020.

[98] Clive H. Bock, Jayme G. A. Barbedo, Anne-Katrin Mahlein, and Emerson M. Del Ponte. A special issue on phytopathometry —

visual assessment, remote sensing, and artificial intelligence in the twenty-first century. *Tropical Plant Pathology*, 47(1):1–4, February 2022.

[99] Jayme Garcia Arnal Barbedo. An Automatic Method to Detect and Measure Leaf Disease Symptoms Using Digital Image Processing. *Plant Disease*, 98(12):1709–1716, December 2014.

[100] Jayme Garcia Arnal Barbedo. Digital image processing techniques for detecting, quantifying and classifying plant diseases. *SpringerPlus*, 2(1):660, December 2013.

[101] C. H. Bock, Poole , G. H., Parker , P. E., and T. R. and Gottwald. Plant Disease Severity Estimated Visually, by Digital Photography and Image Analysis, and by Hyperspectral Imaging. *Critical Reviews in Plant Sciences*, 29(2):59–107, March 2010.

[102] Josafhat Salinas Ruíz, Osval Antonio Montesinos López, Gabriela Hernández Ramírez, and Jose Crossa Hiriart. *Generalized Linear Mixed Models with Applications in Agriculture and Biology*. Springer International Publishing AG, Cham, SWITZERLAND, 2023.

[103] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[104] Feixiang Lu, Haotian Peng, Hongyu Wu, Jun Yang, Xinhang Yang, Ruizhi Cao, Liangjun Zhang, Ruigang Yang, and Bin Zhou. InstanceFusion: Real-time Instance-level 3D Reconstruction Using a Single RGBD Camera. *Computer Graphics Forum*, 39(7):433–445, 2020.

[105] Diana Koldasbayeva, Polina Tregubova, Mikhail Gasanov, Alexey Zaytsev, Anna Petrovskaia, and Evgeny Burnaev. Challenges in data-driven geospatial modeling for environmental research and practice. *Nature Communications*, 15(1):10700, December 2024.

[106] Md Zahangir Alom, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C. Van Essen, Abdul A. S. Awwal, and Vijayan K. Asari. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, 8(3):292, March 2019.

[107] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, June 2011.

[108] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021.

[109] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020.

[110] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui

Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022.

[111] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning, March 2020.

[112] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon Emissions and Large Neural Network Training, April 2021.

[113] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP, June 2019.

[114] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, July 2020.

[115] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

[116] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better?, June 2019.

[117] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN Features off-the-shelf: An Astounding Baseline for Recognition, May 2014.

[118] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu,

Ekin D. Cubuk, and Quoc V. Le. Rethinking Pre-training and Self-training, November 2020.

[119] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020.

Applications Demonstration - Case Studies

This chapter presents three case studies that demonstrate the possibility to get georeferenced observations for different variable types within the EPPO framework to enable geostatistical covariates estimation. Each case study addresses a specific variable type as defined in the EPPO standards.

3.1 Continuous and Discrete Variables: Plant Counting with Machine Learning

Note: The following section is based on published work:

Bumbaca, S.; Borgogno-Mondino, E.C. On the Minimum Dataset Requirements for Fine-Tunining an Object Detector for Arable Crop Plant Counting: A Case Study on Maize Seedlings. Remote Sens. 2025, 17, 2190. DOI: 10.3390/rs17132190

This study investigates minimum dataset requirements for implementing machine learning-based object detection in continuous variable assessment, specifically maize seedling counting in agricultural field trials. The research addresses the critical question of training data sufficiency for deploying computer vision techniques within EPPO standards.

Abstract

Effective object detection in precision agriculture requires understanding minimum dataset requirements, yet this remains undetermined for arable crops seedling detection. This study investigates the minimum dataset size and quality needed to achieve benchmark performance (R^2 = 0.85) across different object detection paradigms. We systematically evaluated many-shot models (YOLOv5, YOLOv8, YOLO11, RT-DETR), few-shot (CD-ViTO), and zero-shot (OWLv2) approaches using orthomosaic imagery of maize seedlings, while also implementing a handcrafted algorithm as baseline. Models were tested with varying dataset sizes, quality levels, and training sources (in-domain vs out-of-distribution). Results demonstrate that no out-of-distribution trained model achieved benchmark performance, while in-domain trained models reached the benchmark with 60-130 annotated images, depending on architecture. Transformer-mixed models (RT-DETR) required fewer samples (60) than CNN-based models (110-130), but showed different sensitivities to annotation quality reduction. Models maintained benchmark performance with 65-90% of original annotation quality. Neither few-shot nor zero-shot approaches met benchmark requirements despite their recent advances. These findings provide practical guidance for efficiently developing maize seedling detection systems, emphasizing that successful deployment requires in-domain training data, with minimum requirements dependent on model architecture.

2.1.1 Introduction

2.1.1.1 Arable Crop Plant Counting by Object Detection

Plant counting is a critical operation in precision agriculture, plant breeding, and agronomical evaluation. Accurate plant detection and counting in agricultural applications serves multiple critical functions:

1. Crop Establishment Assessment: Early-season seedling counts determine whether or not replanting is necessary, directly affecting yield potential and economic returns [1].
2. Precision Agriculture: Individual plant locations enable variable-rate application of inputs (fertilizers, pesticides) and selective harvesting, reducing costs and environmental impact [2].
3. Plant Breeding Programs: Automated counting accelerates phenotyping workflows, enabling evaluation of larger populations and more traits [3].
4. Insurance and Compliance: Standardized counting methods support crop insurance assessments and regulatory compliance for agricultural subsidies [4].
5. Research Applications: Consistent counting methods enable meta-analyses across studies and reproducible research in agronomy [5].

Traditionally performed manually, this labor-intensive task is increasingly automated through computer vision algorithms.

To validate any counting method, establishing a performance benchmark is essential. Such benchmarks can derive from manual counting accuracy, international standards, or comparison with established methods. According to the European Plant Protection Organization, acceptable plant counting methods must achieve a coefficient of determination (R^2) of 0.85 when compared to manual counting [5], corresponding to a Root Mean Square Error (RMSE) of approximately 0.39. This same benchmark ($R^2 = 0.85$) is widely recognized in the scientific literature [6].

In agricultural applications specifically, object detection offers advantages over regression-based counting methods. Studies show object detection not only achieves superior accuracy [6] but also provides geographical plant coordinates when applied to georeferenced orthomosaics, rather than just density estimates. The validation of this ability relies on metrics such as Intersection over Union (IoU), Average Precision (AP), and Average Recall [7] rather than on the coefficient of determination.

Georeferenced orthomosaics—created through aerial photogrammetry from overlapping images with Ground Control Points (GCPs) [8]—are particularly valuable for agricultural counting applications. Their fixed scale and orientation to geographical coordinates simplify object detection by providing consistent object sizes and eliminating perspective distortion. However, georeferenced orthomosaics also present certain limitations. Georeferencing errors due to low-quality

or insufficient GCPs, or reliance on onboard GNSS/IMU systems, can significantly reduce spatial accuracy. This is particularly the case in datasets derived from unmanned aerial vehicles [9, 10]. Moreover, distortions may persist in areas with high relief or vegetation canopy variability if digital elevation models are inadequately detailed [11, 12]. Finally, computational demand and processing time during photogrammetric reconstruction remain a constraint, particularly for high-resolution or large-area orthomosaics [13].

Like many computer vision tasks in specialized domains, agricultural plant counting suffers from data scarcity [14]. Public datasets are limited and often lack critical information like orthorectification parameters or precise scale information. To focus our investigation, we selected grain maize seedlings (*Zea mays* L.) at the V3–V5 growth stage [15] as our case study, as this crop is well represented in both the scientific literature [16, 17] and public repositories [18, 19].

Maize, the world's highest-production crop [20], offers ideal characteristics for object detection at this growth stage. Its regular planting pattern with defined inter-row and intra-row spacing, minimal overlapping, and distinctive appearance makes it suitable for automated counting. These characteristics are shared by other row crops like sunflower (*Helianthus annuus* L.) and sugarbeet (*Beta vulgaris* L.), potentially making our findings applicable to a broader range of agricultural scenarios [21].

2.1.1.2 Evolution of Object Detection Methods

The broader field of object detection has evolved from non-machine learning methods, here named Handcrafted (HC) methods, to sophisticated Deep Learning (DL) approaches. Handcrafted methods rely on explicitly programmed rules and traditional computer vision techniques such as color thresholding, edge detection, and morphological operations to identify objects. These approaches require domain expertise to design feature extraction algorithms but offer interpretability and computational efficiency [16, 22]. While state-of-the-art detection now relies primarily on DL, HC methods still find application in agricultural contexts [16, 23].

Modern DL object detection architectures fall into two main categories: those based on Convolutional Neural Networks (CNNs) [24] like Faster R-CNN [25] and YOLO [26], and those employing Transformer architectures [27] like DETR [28], or hybrid approaches combining both paradigms. The fundamental difference lies in how images are processed—CNNs use grid-based convolutions while Transformers process image patches using attention mechanisms [29].

On standard computer vision benchmarks like COCO [7] and ImageNet [30], Transformer-based approaches generally outperform CNN-based models in accuracy [31]. However, in agricultural applications specifically, CNN-based architectures like YOLO variants remain widely used due to their efficiency with smaller images and lower computational requirements [32, 33]. For agricultural deployment with limited training data, Transformer-based models may offer

advantages in fine-tuning scenarios [34, 35].

For our comparative analysis of agricultural object detection approaches, we selected representative architectures from both paradigms. From the CNN family, we chose YOLOv5 and YOLOv8 due to their widespread agricultural adoption [33]. From Transformer-mixed architectures, we selected RT-DETR [36] and YOLO11 [37], which demonstrate state-of-the-art performance on standard benchmarks.

2.1.1.3 Data-Efficient Detection Methods

A critical challenge in agricultural object detection is the high cost of data annotation. Recently, emerging paradigms like zero-shot and few-shot object detection offer potential solutions by reducing or eliminating annotation requirements. These approaches differ fundamentally from traditional “many-shot” detectors in their data needs.

In few-shot detection, models learn to identify objects from minimal examples—often just 1–30 annotated instances (or “shots”). These methods typically leverage feature transfer or meta-learning [38] to generalize from limited data. The agricultural potential of such approaches is significant, as they could drastically reduce annotation burden for new crop varieties or growth stages.

Zero-shot detection represents the extreme case where models detect novel objects without any labeled examples by exploiting semantic relationships or contextual information learned from other classes [39]. State-of-the-art zero-shot detectors include YOLO-World [40],

OWLv2 [41], and Grounding DINO [42].

While these data-efficient approaches show promise in general computer vision tasks, their effectiveness for agricultural applications remains largely unexplored. For maize seedling detection specifically, only two studies have investigated few-shot methods [43, 44], with neither achieving benchmark performance or clearly specifying shot counts. No studies have yet evaluated zero-shot detection for maize seedling counting.

2.1.1.4 Dataset Requirements for Agricultural Object Detection

While general object detection benchmarks provide standardized evaluation, they poorly represent agricultural conditions. Numerous studies have addressed plant detection in field settings [17, 45, 46], but few have systematically investigated minimum dataset requirements for robust performance [16, 47].

This knowledge gap creates significant challenges for practitioners who must determine how much data to collect and annotate for effective deployment. The agricultural domain's unique characteristics—variable environment, specific plant phenotypes, and orthomosaic imagery—may substantially alter data requirements compared to general computer vision tasks.

General deep learning principles suggest model performance correlates strongly with training data quantity [48] and quality [49]. These relationships can be modeled using empirical approaches [50, 51].

Other factors affecting minimum dataset requirements include model architecture [52], backbone selection [53], and data augmentation strategies [54].

The most critical factor, however, appears to be dataset source. Studies consistently show that using in-domain data (from the same distribution as the inference target) dramatically improves accuracy while reducing required dataset size compared to out-of-distribution training [16, 47]. This finding has profound implications for agricultural deployment scenarios.

2.1.1.5 Study Aim

This study aims to establish the minimum dataset requirements for accurate maize seedling detection in georeferenced orthomosaics across different object detection paradigms. Here, the dataset size and quality are respectively defined as the amount of annotated images in the training set and the accuracy of the annotations. In particular, the annotation quality is defined as the percentage of correct annotations relative to the total ground truth annotations present in each image, with 100% representing perfect annotations where every plant is correctly identified and bounded. For example, if an image contains 10 plants and the annotator correctly identifies and bounds 8 of them, the annotation quality for that image would be 80%. To simulate varying annotation quality levels, we systematically removed a percentage of existing annotations from our complete ground truth dataset, effectively simulating scenarios where human annotators miss a certain proportion of plants due to factors

such as time constraints, fatigue, or challenging visual conditions. This approach allows us to evaluate how robust different models are to incomplete annotations, which is a common real-world scenario in agricultural applications where exhaustive annotation can be prohibitively expensive or time-consuming. Specifically, we investigate the following:

1. How training data source (in-domain vs. out-of-distribution) affects required dataset size
2. Minimum dataset size needed to achieve benchmark performance ($R^2 = 0.85$) for different model architectures
3. Minimum annotation quality required to maintain benchmark performance
4. Whether or not newer few-shot and zero-shot approaches can meet agricultural performance standards with reduced annotation requirements
5. The potential role of handcrafted methods in modern deep learning pipelines

By systematically varying training dataset size (10–150 images), annotation quality (10–100%), and evaluating diverse architectures (CNN-based, Transformer-based, few-shot, and zero-shot), we provide comprehensive guidance for implementing efficient maize seedling detection systems. Our findings establish empirical relationships between dataset characteristics and model performance, offering prac-

tical insights for optimizing the annotation effort versus detection performance trade-off in agricultural object detection.

2.1.2 Materials and Methods

2.1.2.1 Datasets

The datasets used in this study to train the object detection models for maize seedling counting are nadiral or supposedly nadiral images of maize seedlings at the V3–V5 growth stage, or estimated so. The V3–V5 growth stage is defined by the BBCH scale as the stage where the third to fifth leaf is unfolded and the plant is 15–30 cm tall [15].

This study uses two dataset sources as training sets: the Out-of-Distribution (OOD) dataset and the In-Domain (ID) dataset. The ID datasets are from the same source as the testing dataset, while the OOD datasets are not. The OOD datasets are composed of images from scientific literature [16, 55] and from internet repositories [18, 19]. The ID datasets were collected during this study. This ID dataset creation consisted of capturing nadiral images of three study areas with a Phantom 4 Pro v2.0 (DJI, Shenzhen, China) drone equipped with its default series RGB camera at about 10 m above ground for a Ground Sampling Distance (GSD) of 2.7 mm/pixel. The number of images captured depends on the study area size, which was about 2 hectares for the ID_1 location and about 1 ha for the other two. For each location, an orthomosaic was created using photogrammetric software. Bundle ad-

justment error was estimated as 38 mm using the GCPs surveyed by GNSS operating in VRS-NRTK mode. The orthomosaics were generated with an average GSD of 5 mm/pixel in the WGS84/UTM 32 N reference system. We chose this GSD because it is the minimum GSD that allows the detection of maize seedlings at the V3–V5 growth stage with a nadiral camera [56].

The OOD scientific datasets consist of tiles of georeferenced orthomosaics of maize seedlings from the scientific literature. The OOD internet datasets consist of RGB images of maize seedlings from internet repositories. The ID datasets were collected during this study and consist of tiles of georeferenced orthomosaics of maize seedlings of known scale. The OOD scientific datasets and the ID datasets are composed of tiles of georeferenced orthomosaics of known scale, while the OOD internet datasets are simple RGB images of unknown scale. All the OOD datasets came with annotations, while the ID datasets were manually annotated. The OOD dataset annotations are rectangular bounding boxes centering on an individual plant stem. ID dataset annotation was done during this study by an agronomist by observing the entire orthomosaic in a Geographical Information System (GIS) environment, with the tile grid overlapping the orthomosaic to focus on target tiles without losing the surrounding context, so without losing bordering plants. Annotations were created as squared bounding boxes of a size length equal to the minimum distance between two plants in the row, with each box centered on an individual seedling stem.

Table 1 summarizes the datasets used in this study.

Table 1: Summary of datasets used in the study.

Dataset	Phenological Stage	Train Size	Test Size
OOD Scientific			
DavidEtAl.2021 [16]	V3	182 tiles	N/A*
LiuEtAl.2022 [55]	V3	596 tiles	N/A*
OOD Internet			
OOD_int_1 [18]	V3	216 tiles	N/A*
OOD_int_2 [19]	V5	174 tiles	N/A*
ID [57]			
ID_1	V3	150 tiles	20 tiles
ID_2	V3	150 tiles	20 tiles
ID_3	V5	150 tiles	20 tiles

* N/A indicates that these datasets were used only for training purposes and do not have separate test sets in this study.

To make the two types of datasets comparable, we chose to rescale the images to a scale of 0.005 m/pixel where the scale was known (scientific OOD and ID datasets), obtaining orthomosaics of different

sizes. All the orthomosaics were then cropped to 224 $\times$ 224 pixel tiles. This tile size was selected because at 5 mm/pixel resolution it covers 1.12 × 1.12 m of field area. Given that typical grain maize inter-row distance is 0.75 m, this size enables capturing approximately two rows per tile, which is optimal for row pattern identification in the HC algorithm and provides sufficient context for object detection models [16, 55]. This particular image size was also chosen as a standard from AlexNet [58] as it should be compatible with most of the object detection architectures. The annotations were rescaled and cropped where needed. Figure 1 shows a sample for each dataset. Each ID dataset has 20 tiles to be used as the testing dataset, while the other 150 tiles are used as the training dataset.

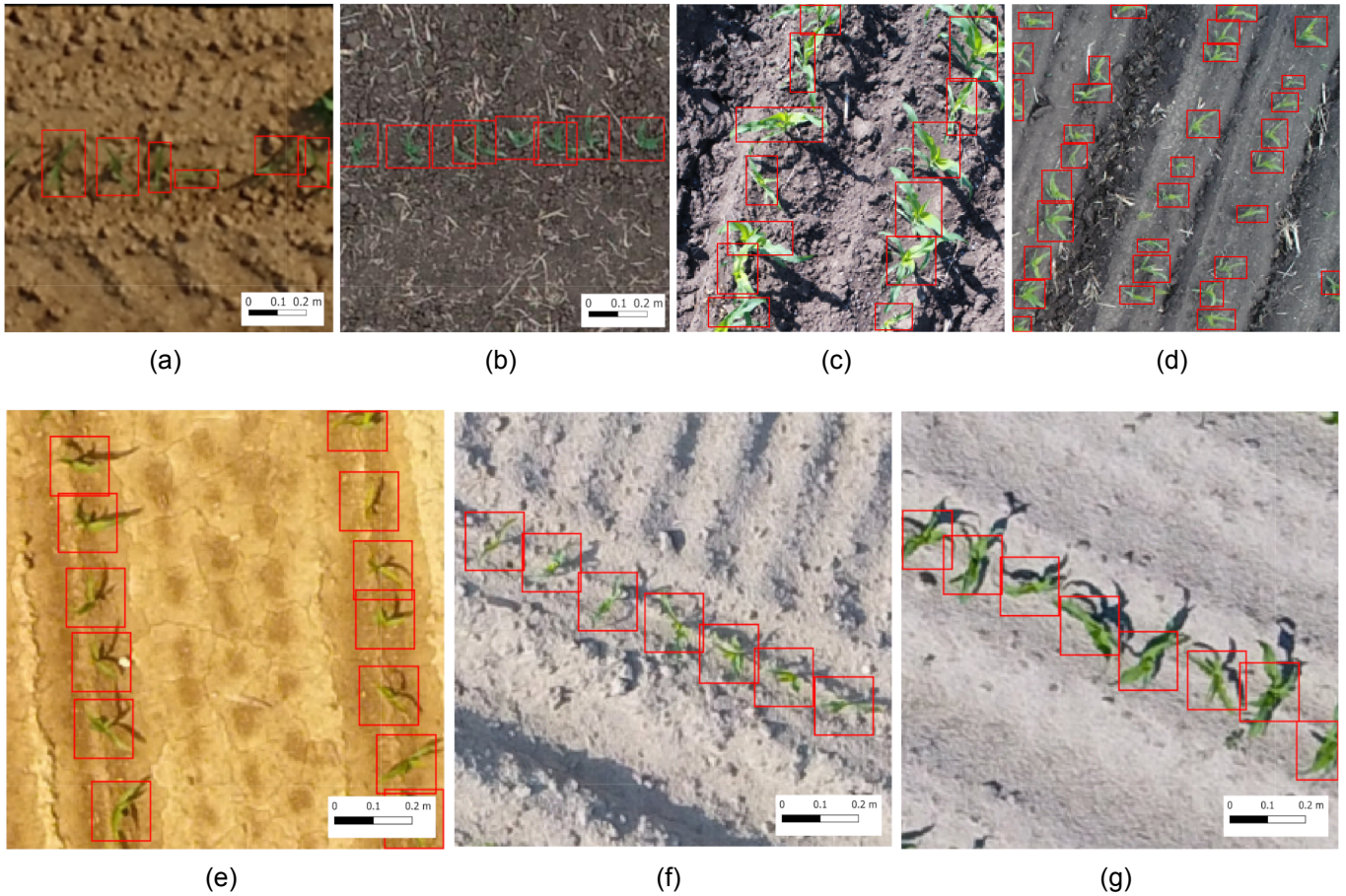

Figure 1: Image examples taken from each dataset, ground truth bounding boxes are shown in red. (**a**) DavidE-tAl.2021, (**b**) LiuEtAl.2022, (**c**) Internet Maize stage V3, (**d**) Internet Maize stage V5, (**e**) ID_1, (**f**) ID_2, (**g**) ID_3.

2.1.2.2 Handcrafted Object Detector

Like other works [16, 23, 55], we wrote an HC algorithm to obtain annotated tiles from the orthomosaics, basing it on agronomical knowledge and color thresholding. Hue, Saturation, and Value (HSV) color space was used here to threshold the image, to obtain green pixels, but other color spaces can be used. For the execution of this algorithm, the following graphical and agronomical parameters must be set: color minimum and maximum thresholds (color threshold), the minimum and maximum leaf area for the plant (leaf area range), the minimum distance between plants on rows (intra-row distance), and the distance between rows (inter-rows distance). The algorithm is expected to work on the orthomosaics of maize seedlings at the V3–V5 growth stage, with low weed infestation, with rows having roughly the same angle with meridian and distance between them. The algorithm was implemented in Python 3.13.1 using the following packages: numpy (v 1.24.3), torch (v 2.0.1), PyYAML (v 6.0), rasterio (v 1.3.8), shapely (v 2.0.1), fiona (v 1.9.4), scikit-image (v 0.21.0), scikit-learn (v 1.3.0), matplotlib (v 3.7.1). The complete implementation is accessible from `https://gist.github.com/SamueleBumbaca/4a227bbe7b78d6be3424899c16c60bb4` (accessed on 20 June 2025).

The algorithm is divided in two sequential parts that form a detection–verification pipeline. The first part, named HC1 algorithm 1, performs initial plant detection by thresholding pixels within the specified color range, identifying connected regions, and filtering them based on expected leaf area. HC1 outputs region polygons representing potential plants, but typically includes many false positives due to its

simple color-based approach. To address this limitation, we implemented a second process named HC2 algorithm 2 that applies agronomical knowledge of field structure. HC2 filters the HC1 output by verifying that detected plants form proper row patterns with expected intra-row and inter-row spacing. It uses RANSAC [59] to identify the linear alignments of plants and validates that these alignments match expected field geometry (consistent row slope and spacing). Only tiles where HC2 confirms the expected number and arrangement of plants are retained for the final dataset. This two-stage approach enables the automated extraction of high-confidence annotations from the orthomosaics. Complete algorithm pseudocodes are provided in the Appendix A.

2.1.2.3 Deep Learning Object Detectors

Many-shot model selection criteria: From the extensive landscape of available object detection architectures, we applied the following selection criteria: (1) widespread adoption in agricultural computer vision applications [33], (2) availability of multiple model sizes to investigate parameter count effects on dataset requirements, (3) representation of different architectural paradigms (CNN-based vs. Transformer-based), (4) consistent implementation framework for fair comparison, and (5) proven performance on small object detection tasks relevant to seedling identification. The many-shot object detectors used in this study include YOLOv5, YOLOv8, RT-DETR, and YOLO11. Each represents distinct architectural approaches to object detection:

- **YOLOv5** with model descriptions uses a CSP (Cross Stage Partial) backbone with a PANet neck for feature aggregation, achieving efficient detection through grid-based prediction with multiple anchors per cell. It was selected as the baseline CNN architecture due to its dominance in agricultural applications [33] and extensive use in crop monitoring studies [45, 46]. Its CSP backbone with PANet neck provides a well-established reference point for minimum dataset requirements in agricultural contexts.

- **YOLOv8** with model descriptions improves upon YOLOv5 by adopting a more efficient C2f block in its backbone, implementing an anchor-free detection head and using a task-specific decoupled head design for better accuracy–speed trade-offs. It represents the next-generation YOLO evolution with anchor-free detection and improved C2f blocks. We included it to evaluate whether or not architectural improvements could reduce dataset requirements compared to YOLOv5, given its reported superior accuracy–efficiency trade-offs [60].

- **YOLO11** with model descriptions further refines the architecture with multi-scale deformable attention, improving small object detection—a crucial feature for seedling identification in varied field conditions. It was chosen as the most recent YOLO variant including an attention mechanisms (transformer). This selection allows us to assess whether or not state-of-the-art Transformer-mixed improvements affect minimum dataset requirements for small object detection.

- **RT-DETR** with model descriptions represents a hybrid approach combining CNN backbones with Transformer decoders, utilizing deformable attention for adaptive feature sampling and parallel prediction heads for real-time performance. Unlike pure CNN-based YOLO variants, RT-DETR's Transformer components enable the modeling of global relationships between objects. We selected it over pure Transformer models (like DETR) due to its real-time capabilities and proven performance on agricultural datasets [36]. Its inclusion allows direct comparison between CNN-only and hybrid approaches for minimum dataset requirements.

Excluded alternatives and rationale: with model descriptions We deliberately excluded several model families: Faster R-CNN and other two-stage detectors due to their computational overhead and limited agricultural adoption; pure Transformer models like DETR due to prohibitive training requirements for small datasets [28]

We used the Ultralytics implementation for all models as it is open-source [61] and enables consistent parameter tuning across architectures.

All model training and inference were performed on a workstation equipped with an Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30 GHz, 64.0 GB RAM, and an NVIDIA RTX A5000 GPU with 24 GB VRAM. The computational constraints influenced certain experimental design choices, such as batch size and precision settings.

For all many-shot models we used the same hyperparameters and augmentations as the library default, with the following exceptions:

- batch size: 16 (increased from default 8 to maximize GPU utilization while maintaining stable gradients)
- maximum training epochs: 200 (extended from default 100 to ensure convergence with small datasets)
- maximum training epochs without improvement: 15 (increased from default 10 for early stopping to allow longer plateau exploration)
- precision: mixed (to balance training speed and numerical accuracy)

The default augmentations from the Ultralytics library include random scaling ($\pm$10%), random translation ($\pm$10%), random horizontal flip (probability 0.5), HSV color space augmentation (hue $\pm$0.015, saturation $\pm$0.7, value $\pm$0.4), and mosaic augmentation. These augmentations were selected to reflect potential variations in field conditions without introducing unrealistic distortions.

For few-shot detection, we employed CD-ViTO, which differs fundamentally from many-shot approaches:

- **CD-ViTO** with model descriptions uses cross-domain prototype matching, where a small set of annotated examples (shots) serves as class prototypes. It leverages Vision Transformer (ViT) backbones to extract feature representations and computes similarity between query images and prototypes for object localization. This architecture is specifically designed for scenarios with limited training data.

The size of this model is determined by the backbone used: ViT-S, ViT-B, or ViT-L [62]. We used the implementation provided by the authors [63]. In our study, a 'shot' corresponds to an image with a single annotated plant. We tested 1, 5, 10, 30, and 50 shots, randomly selected from the ID manually labeled dataset.

For zero-shot detection, we selected OWLv2:

- **OWLv2** with model descriptions represents a fundamentally different paradigm that requires no labeled examples of the target class. It leverages large-scale pre-training on image–text pairs to establish connections between visual features and natural language descriptions. During inference, it detects objects based solely on text prompts, eliminating the need for class-specific training data entirely.

We chose OWLv2 as our zero-shot exemplar because it represents state-of-the-art performance in open-vocabulary detection [41, 42]. For testing, we used the implementation from the Transformers library [64] with the published parameters. We evaluated two encoder sizes (ViT-B/16, ViT-L/14) with three pre-training strategies:

- **Base models**: Trained using self-supervised learning with the OWL-ST method, generating pseudo-box annotations from web-scale image–text datasets.
- **Fine-tuned models**: Further trained on human-annotated object detection datasets.

- **Ensemble models**: Combining multiple weight-trained versions to balance open-vocabulary generalization and task-specific performance.

For all OWLv2 variants, we tested multiple text prompts to describe maize seedlings, ranging from simple terms (“maize”, “seedling”) to more descriptive phrases (“aerial view of maize seedlings”, “corn seedlings in rows”). The complete list of eleven prompts is provided in the Appendix A.

The choice of text prompt significantly influences OWLv2 performance, as the model relies on semantic alignment between visual features and language descriptions learned during pre-training [65]. Simple generic terms like “maize” or “plant” may activate broader visual concepts that include mature plants or other crop types, potentially reducing detection specificity. Conversely, descriptive phrases like “aerial view of maize seedlings” provide more contextual information that should theoretically improve alignment with our orthomosaic imagery [66]. However, if such specific descriptions were underrepresented in the pre-training data, they may perform worse than simpler terms [67]. To account for this variability, we evaluated all prompts systematically and report the results using the best-performing prompt for each model variant; Figure 9 shows the full distribution of performance across all tested prompts.

table 2 shows the architectures used in the study with their parameter size specifications.

Table 2: Summary of tested architectures and model sizes.

Architecture	Shots	n * or S	s * or S	m * or B	l * or L	x *
YOLOv5	many	1.9	7.2	21.2	46.5	86.7
YOLOv8	many	3.2	11.2	25.9	43.7	68.2
YOLO11	many	4.0	12.5	28.0	50.0	75.0
RT-DETR	many	-	-	-	60.0	80.0
CD-VITO	few	-	22.0 †	86.0 ‡	307.0 §	-
OWLv2	zero	-	-	86.0 ‡	307.0 §	-

Values represent millions of parameters. * Model size variants stand for nano (n), small (s), medium (m), large (l), and extra-large (x). † ViT-S (Small) backbone. ‡ ViT-B (Base) backbone. § ViT-L (Large) backbone.

2.1.2.4 Minimum Dataset Size and Quality Modelling

In order to investigate the minimum size and quality of the dataset required to train a robust object detection model for maize seedling counting, we conducted a series of experiments where the above mentioned DL models were recursivelly fitted with increasing dataset size and quality. To evaluate the HC object detector on the ID dataset, we measured three key aspects: (1) the number of tiles that the HC algorithm successfully annotated from the orthomosaics (tiles), (2) this number as a percentage of the total available dataset (dataset%), and (3) the detection accuracy compared to manual annotations us-

ing standard metrics (R^2, $RMSE$, $MAPE$, and mAP). For many-shot models we consider a training dataset split of 10% validation and 90% training, while for few-shot the number of shots determined the amount of training samples. Zero-shot learning relies only on descriptions of the objects to be detected in natural language. Thus, the effects of the prompts were evaluated. As previously mentioned, this involved testing multiple text prompts. For what concerns only the dataset size evaluation, for many-shot models we considered sizes from 10 to 150 images in 15 steps of 10 images, while for few-shot models we considered 1, 5, 10, 30, and 50 shots. As concerns the dataset quality, we evaluated the annotation quality by reducing the number of annotations per image from 100% to 10% in 10 steps of 10% while keeping the dataset size constant. To evaluate the influence of dataset source (OOD or ID) on the model performance, we trained the models using both OOD and ID datasets with the same experimental protocol. Only the most relevant results are reported.

For all the models, we evaluated the relationship between dataset size or quality and model performance using R^2 and mAP, respectively, for plant counting and plant detection. Whether or not R^2 provided values below -1 [68], we also considered $RMSE$ as the metric for counting [69]. $MAPE$ was considered for few-shot and zero-shot models only to evaluate the quality of the annotations produced by the prediction of these models. We list here the metrics formulas for clarity:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

where y_i is the ground truth count for the i-th image, $\hat{y}_i$ is the predicted count, and $\bar{y}$ is the mean of all ground truth counts. R^2 ranges from $-\infty$ to 1, with 1 indicating perfect prediction, 0 indicating that the model predictions are no better than simply predicting the mean, and negative values indicating that the model performs worse than predicting the mean [68, 69].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

where $RMSE$ measures the average magnitude of prediction errors in the original units (number of plants) for each image [69].

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

where $MAPE$ measures the percentage error relative to the actual values, providing a scale-independent measure of accuracy. It is expressed as a percentage, with lower values indicating a lower percentage of false positives or false negatives. Thus, it was reported as an index of the quality of the annotations. Note that $MAPE$ is only calculated for cases where $y_i \neq 0$ to avoid division by zero. It is particular useful for counting as testing tiles never have zero plants [70].

For object detection performance, we used the standard COCO evaluation metric [7, 71]:

$$mAP = \frac{1}{|IoU|}\sum_{t \in IoU} \mathrm{AP}_t \tag{4}$$

where mAP (mean Average Precision) is calculated at a single IoU (Intersection over Union) threshold of 0.5. AP at the IoU threshold is the area under the precision–recall curve for detections that meet that IoU threshold criterion.

To test the predictability minimum dataset size and quality required to train a robust (achieving benchmark) object detector for maize seedling counting through empirical models, we test the logarithmic, arctan, and algebraic root functions to fit the dataset size or quality versus performance relationships, as suggested by previous studies [51].

These functions were selected because they represent different theoretical behaviors commonly observed in machine learning scaling studies:

Logarithmic function: Models the diminishing returns pattern where initial data additions provide substantial performance gains, but additional data yield progressively smaller improvements. This behavior is theoretically grounded in learning theory, where models approach their optimal performance asymptotically [50].

Arctangent function: Represents a saturating behavior where performance increases rapidly at first, then approaches a plateau. This function is particularly suitable for modeling performance metrics bounded by theoretical limits (e.g., R^2 approaching 1.0) and captures scenarios where models reach their maximum capacity given architectural constraints [72].

Algebraic root function: Models power-law relationships between dataset size and performance, allowing for various growth rates de-

pending on the exponent. This function can capture both sub-linear and super-linear scaling behaviors, providing flexibility for different model architectures and learning dynamics [50].

For clarity, we list here the functions tested:

$$\text{Logarithmic: } f(x) = a\ln(x) + b \tag{5}$$

$$\text{Arctan: } f(x) = a\arctan(bx) + c \tag{6}$$

$$\text{Algebraic Root: } f(x) = ax^{1/b} + c \tag{7}$$

where x represents dataset size (number of images) or quality (percentage of correct annotations) and a, b, and c are fitted parameters that determine the shape and scale of each function.

For each model architecture and metric combination, we fitted all three functions to the observed dataset size versus performance data points using least squares regression. The function yielding the highest goodness-of-fit was selected as the best predictor for that specific model–metric combination. This approach allows us to identify which scaling pattern best describes each model's behavior and enables prediction of the minimum dataset size required to achieve benchmark performance.

The selected empirical model can then be used to interpolate or extrapolate performance estimates for untested dataset sizes, providing practical guidance for annotation planning. For instance, if a log-

arithmic function best fits the data, practitioners can expect diminishing returns from additional annotations beyond a certain point. Conversely, if an algebraic root function provides the best fit, the scaling behavior may indicate more linear or super-linear returns, suggesting different annotation strategies.

For the model fits to dataset size versus performance relationships, we evaluated multiple fitting functions and selected the one with the highest goodness-of-fit:

$$GoF = R^2_{\text{fit}} = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

where y_i is the observed metric (either R^2 or mAP), $f(x_i)$ is the fitted value at dataset size x_i, and $\bar{y}$ is the mean of the observed metrics.

All the trained models were tested on the testing dataset tiles with the SAHI method [73]. SAHI (Slicing Aided Hyper Inference) is a technique designed to improve object detection performance on high-resolution images by addressing the scale mismatch between training and inference conditions. The method slices the testing image into smaller overlapping segments (patches) of the same size as the training tiles (224 × 224 pixels in our case) and then applies the model to each patch independently. The model outputs from each patch are then merged using non-maximum suppression to eliminate duplicate detections, and the final results are cropped to the original tile boundaries.

The use of SAHI is justified in our context because while models are trained on 224 × 224 pixel tiles, the real-world application in-

volves inference on larger orthomosaics where objects may be partially occluded by tile boundaries or appear at different scales. SAHI overcomes this limitation by ensuring that all potential objects are evaluated by the model as complete entities rather than fragmented across tile boundaries. This approach is expected to provide better performance compared to using single tiles as input, as it reduces boundary effects and maintains consistent object scale during inference.

The predictions were then thresholded by a list of confidence score thresholds to obtain the plant count. All the metrics were computed for different score thresholds for all the models to evaluate the model performance at different confidence levels. The values to thresholds bounding boxes score were 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.29, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99. The highest R^2 value within the thresholds was considered as the model performance for that experiment.

2.1.3 Results

2.1.3.1 Handcrafted Object Detector

Table 3 shows the performance of the HC object detector on the ID datasets by enumerating metrics and successfully annotated tiles. The metrics were computed on the testing dataset tiles.

Table 3: HC object detector performance.

Dataset	R^2	$RMSE$	$MAPE$	mAP	Tiles	Dataset %
ID_1	0.95	0.12	9%	0.87	1184	7.8%
ID_2	0.93	0.11	12%	0.81	279	4.2%
ID_3	0.87	0.18	16%	0.73	158	1.8%

The HC algorithm was able to extract a discrete amount of annotated tiles from the orthomosaics, with a percentage of the dataset ranging from 1.8% to 7.8%. Overall, the HC object detector performed well on this set, with R^2 values above 0.85 for all the datasets. The $RMSE$ values were below 0.2, while the mAP values were above 0.7. The MAPE values were below 20% for all the datasets.

In nominal scale, the number of tiles successfully annotated by the HC algorithm was not constant, but always over 150 tiles.

2.1.3.2 Many-Shot Object Detectors

OOD Training

The OOD scientific datasets “DavidEtAl.2021” and “LiuEtAl.2022” were tested singularly and in combination in the experiment named “scientific OOD”. The OOD internet datasets “internet OOD” were tested singularly and in combination with the OOD scientific datasets in the experiment named “All OOD”. Each model and OOD dataset combination was tested on the testing dataset tiles of the three ID datasets.

None of the dataset combinations reached the benchmark R^2 value of 0.85 with any model. The coefficients of determination and the root mean square errors for all the OOD experiments are shown in Figure 2. The Goodness-of-Fit (GoF) values for the R^2 values were always low (below 0.2) for all the metrics. The lowest $MAPE$ value was slightly less than 20%. For these same models, the mAP values were the highest, with the best model being YOLOv8n with the LiuEtAl.2022 dataset. No particular model size seems to provide better results with respect to the others, and neither does the increasing dataset size seem to drive a model size performance trend. As no model achieved the benchmark, no study was done on the dataset quality requirements to achieve such a benchmark.

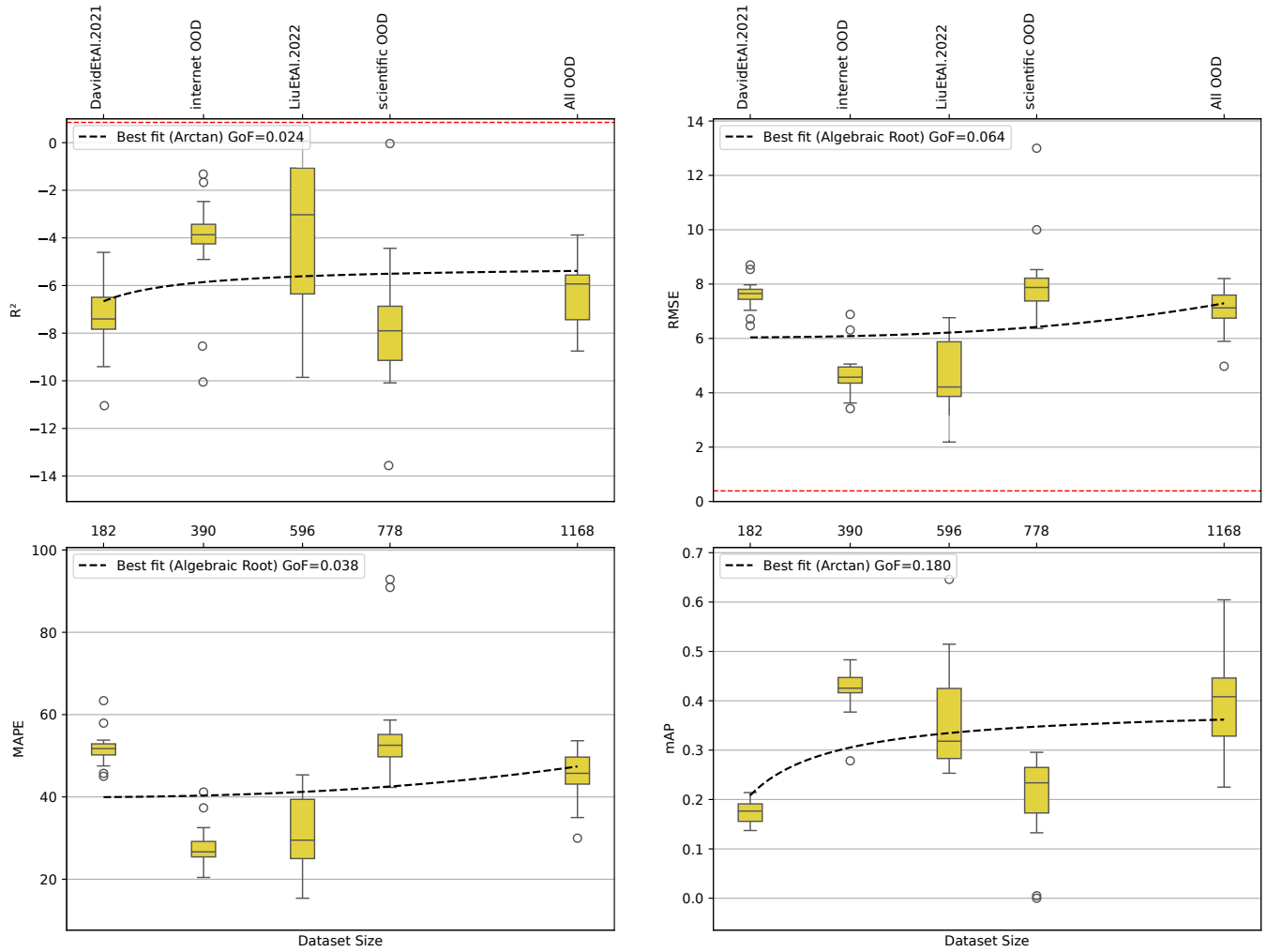

Figure 2: The figure shows the performance of the many-shot object detection models trained on the different Out-of-Distribution (OOD) datasets. The subplots represent four different metrics: R^2, $RMSE$, MAPE, and mAP, respectively, at the right top, left top, left bottom, and right bottom. Each subplot contains the boxplots positioned at the corresponding dataset size values and indicating the distribution of all the model prediction metric values for each dataset. Benchmark thresholds are indicated with red dashed horizontal lines for R^2 (0.85) and $RMSE$ (0.39). Best fit lines for each metric are plotted using different fitting functions, indicated with black dashed lines. GoF values and best model are shown in the legend. A secondary x-axis at the top of each subplot shows the dataset names corresponding to the dataset sizes.

ID Training

The relationship between ID training dataset size and model performance was evaluated for all model architectures and sizes, as shown in Figures 3 and 4. The dataset quality was tested later, taking the combination of model architecture, model size, and training dataset size that achieved the benchmark and retraining that model while reducing the amount of annotations for each tile. The R^2 values of the counting and the mAP values for all models were regressed against the dataset size using a logarithmic, root, or arctan model. The best fitting within them was selected for each model and metric and the GoF was calculated. A high GoF value indicates that model performance is highly predictable by dataset size. Conversely, a poor GoF could indicate that other variables play a more important role in determining model performance, or that the chosen dataset size interval is too narrow to achieve a good fit.

For the combinations of model-architecture/dataset-size that achieved the benchmark, the minimum dataset quality required to achieve the benchmark was evaluated, as shown in Figure 5. The minimum dataset quality was determined by identifying the quality percentage where both the empirical model prediction and the entire confidence interval of the performance metrics remained above the benchmark threshold.

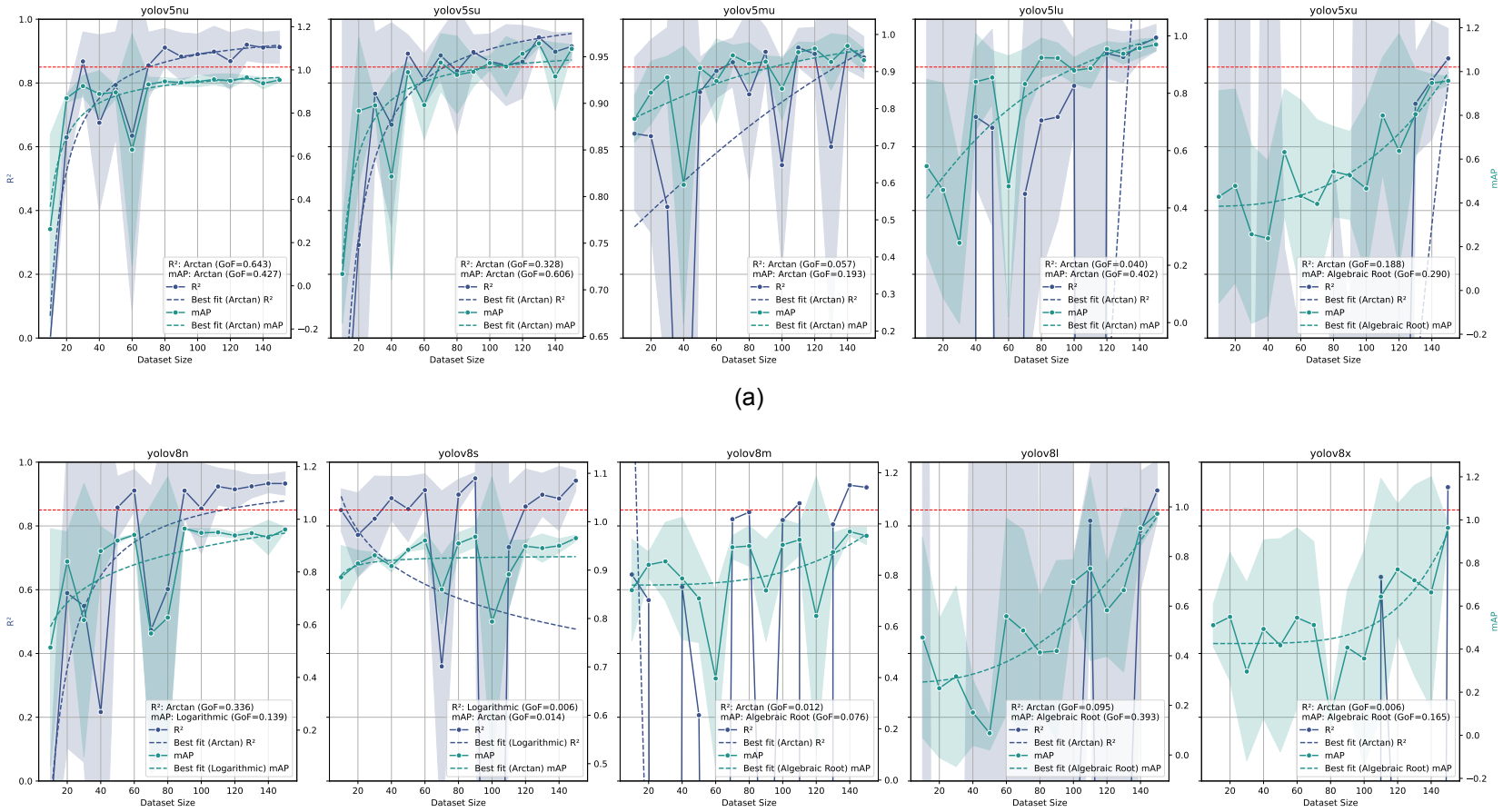

Figure 3: Relationship between dataset size and model performance for CNN-based object detection models (YOLOv5 (**a**) and YOLOv8 (**b**)) trained and tested on ID datasets. On the same line, each subplot represents a different parameters size of the model, increasing from the left to the right. The x-axis represents the dataset size, while the left and right y-axis represents the R^2 and mAP values, respectively. The solid lines represent the mean values, while the dashed lines indicate the logarithmic fit. The shaded area around the solid lines represents the confidence interval (standard deviation) of R^2 or mAP. The red dashed horizontal line represents the benchmark R^2 value of 0.85. The combined legend in the lower right corner of each subplot shows the Goodness of Fit (GoF) for both R^2 and mAP.

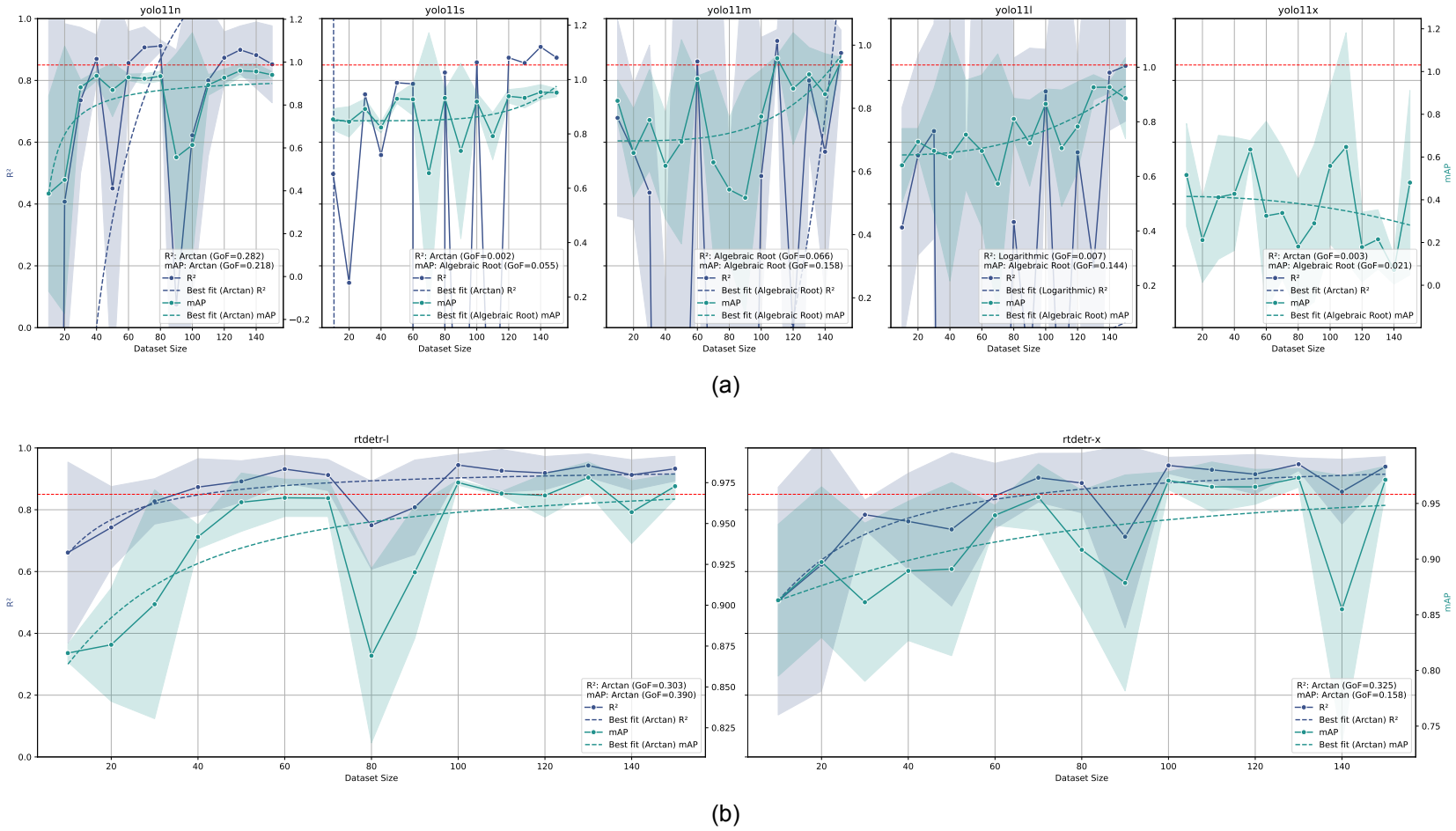

Figure 4: Relationship between dataset size and model performance for Transformer-mixed object detection models (YOLO11 (**a**) and RT-DETR (**b**)) trained and tested on ID datasets. On the same line, each subplot represents a different parameters size of the model, increasing from the left to the right. The x-axis represents the dataset size, while the left and right y-axis represents the R^2 and mAP values, respectively. The solid lines represent the mean values, while the dashed lines indicate the logarithmic fit. The shaded area around the solid lines represents the confidence interval (standard deviation) of R^2 or mAP. The red dashed horizontal line represents the benchmark R^2 value of 0.85. The combined legend in the lower right corner of each subplot shows the Goodness of Fit (GoF) for both R^2 and mAP.

Within YOLO models, YOLOv5n, YOLOv5s, and YOLOv8n achieve the benchmark R^2 value of 0.85 with 130, 130, and 110 samples, respectively, considering the dataset sizes where all three model performances were above 0.85 R^2 and the logarithmic model predicted over-benchmark values for that dataset size. RT-DETR L and RT-DETR X achieve the benchmark R^2 value of 0.85 with 60 and 100 samples, respectively, with the same assumptions as for the YOLO models. For these models, the GoF was above 0.3, while for the models that did not reach the benchmark R^2 value the GoF was always below this value. The mAP seems to follow the same trend as the R^2 values. All the models show a clear trend of increasing R^2 and mAP values as the dataset size increases, as expected. It is also clear that increasing the number of parameters and model complexity for mostly CNN-like models (YOLOs) leads to increasing need for dataset size. For the mostly transformer-like models (RT-DETRs), it is not that clear, also because of the low amount of model parameter sizes tested. The confidence interval reduction as a function of the dataset size indicates that variability in performance decreases significantly as dataset size increases for all models. Taking into account the dataset quality in the same way as done for the dataset size, both quality tests and quality models achieved the benchmark, with 85%, 90%, 85%, and 65% of the original dataset quality for YOLOv5n, YOLOv5s, YOLOv8n, and RT-DETR X, respectively. RT-DETR L did not achieve the benchmark for any dataset quality reduction tested. Overall, RT-DETR performed best in terms of both dataset size and quality. The best performance in terms of dataset size was achieved with RT-DETR trained on 60 images (see predic-

tion examples in Figure 6), while the best performance in terms of dataset quality was achieved with RT-DETR trained on 100 images with a 35% reduction in quality (see Figure 7 for prediction examples).

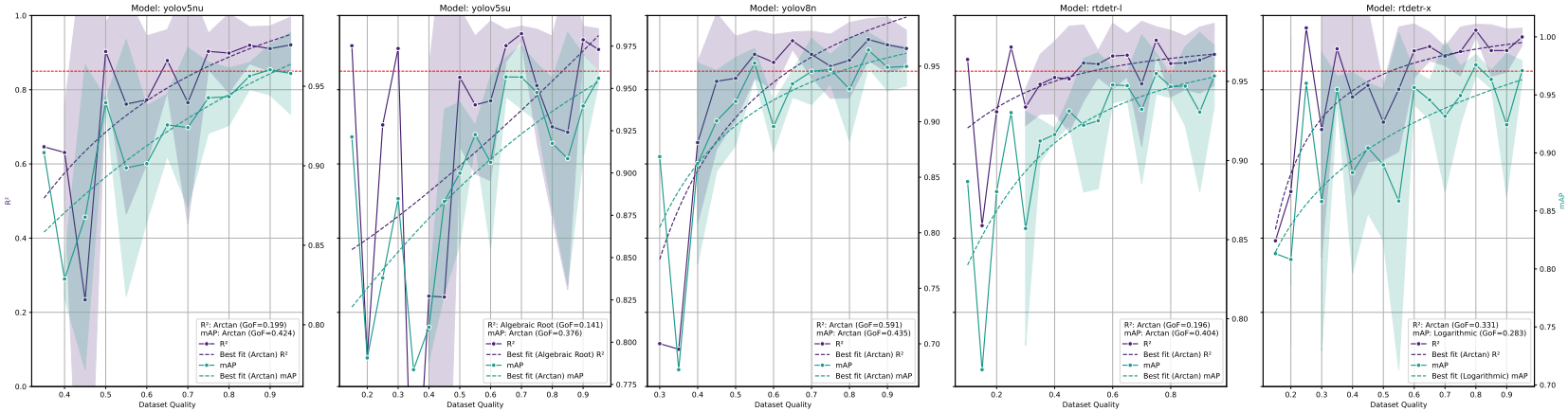

Figure 5: Relationship between dataset quality and model performance for all object detection models that achieved the benchmark. The x-axis represents the dataset quality, while the left y-axis represents the R^2 values. The red dashed horizontal line represents the benchmark R^2 value of 0.85. The legend in the lower right corner of the subplot shows the Goodness of Fit (GoF) for R^2.

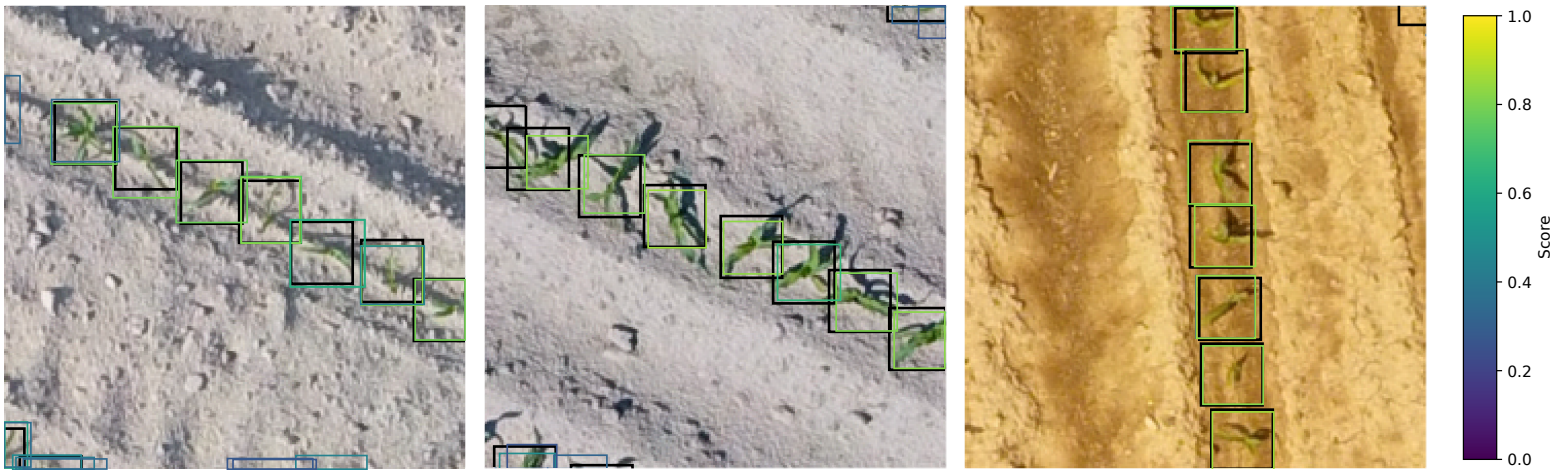

Figure 6: Predictions from the RT-DETR L trained on 60 images. From the left hand side to the right hand side, the images show the 1, 2, and 3 ID test dataset tile examples. The predicted bounding boxes in the images are the ones before non-maximum suppression and threshold. Black bounding boxes are the ground truth annotations, while the bounding boxes in the viridis color scale are the model predictions.

Figure 7: Predictions from the RT-DETR X trained on 100 images with a reduction in quality of 35%. From the left hand side to the right hand side, the images show the 1, 2, and 3 ID test dataset tile examples. The predicted bounding boxes in the images are the ones before non-maximum suppression and threshold. Black bounding boxes are the ground truth annotations, while the bounding boxes in the viridis color scale are the model predictions.

2.1.3.3 Few-Shot Object Detectors

The few-shot models were evaluated against the established benchmarks (R^2 of 0.85 and $RMSE$ of 0.39) using the metrics R^2 and $RMSE$; however, none of the models reached these benchmarks.

The best result achieved by the CD-ViTO model was an $RMSE$ of 3.9 with ViT-B backbone and 50 shots to build the prototypes, which is substantially worse than the benchmark value of 0.39 (10 times higher). This corresponds to a $MAPE$ on counting of about 25% and a mAP of about 0.5, as shown in Figure 8. It corresponds roughly to a miscounted plant over four as it is visible, looking to some predictions of this model in Figure 9. The models fitted on metrics show a reliable GoF for all the metrics, indicating that the model performance is highly predictable by the number of shots. These also show that any CD-ViTO size model would not achieve the benchmark with any shot amount, even if the number of shots were increased beyond those tested.

2.1.3.4 Zero-Shot Object Detectors

Figure 10 shows the relationship between the zero-shot model settings and model performance tested on ID testing datasets. Not all the model settings were able to predict the whole testing dataset. For example, the owlv2-base-patch16-finetuned model was not able to generate any prediction with any prompt for any image of the ID testing. A dataset size relationship with metrics could not be established because zero-shot models do not require fine-tuning training

data. None of the zero-shot model settings reached the benchmark. This is particularly true for the R^2 values, which were always below 0, indicating poor predictive performance. The $RMSE$ values ranged from approximately 5 to 25, significantly higher than those observed in the many-shot and few-shot models. Additionally, $MAPE$ values were also considerably elevated, ranging from around 40 to 140. Furthermore, the mAP values were lower than those of the many and few-shot models for all model settings, except for the owlv2-large-patch14-finetuned model, for which very few images were successfully predicted with an mAP comparable to that of the best few-shot model (50-shots ViT-B backbone). Some rare cases of good predictions were even more accurate than the few-shot best performance, as shown in Figure 11.

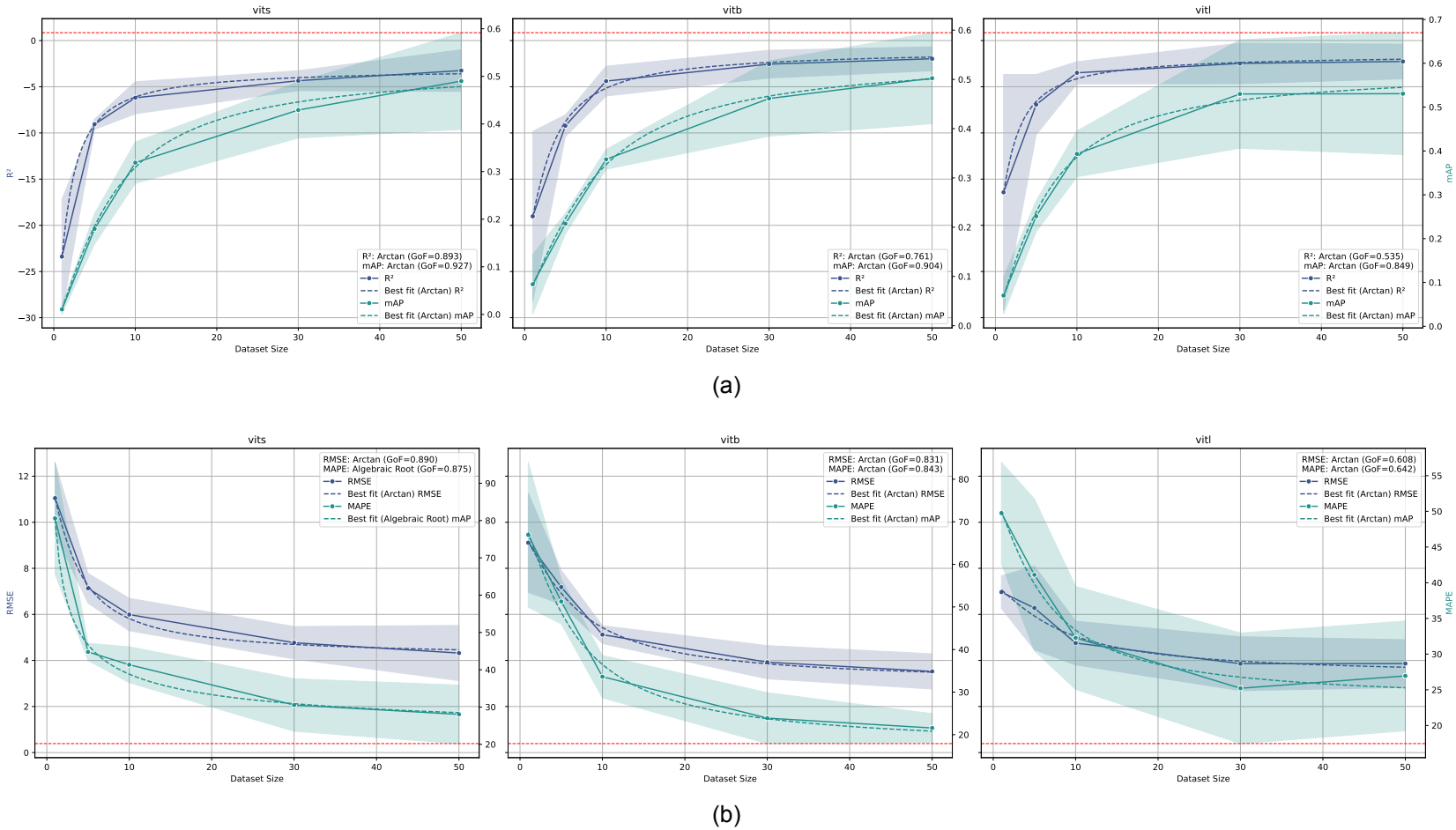

Figure 8: The figure shows the relationship between shots and model performance for the CD-ViTO model trained and tested on ID datasets. The x-axis represents the number of shots. The solid lines represent the mean values, while the dashed lines indicate the shot amount/metric prediction model. The shaded area around the solid lines represents the confidence interval (standard deviation) of the metric. (**a**) The left and right y-axis represents the R^2 and mAP values, respectively. The red dashed horizontal line represents the benchmark R^2 value of 0.85. The combined legend in the lower right corner of each subplot shows the Goodness of Fit (GoF) for both R^2 and mAP. (**b**) The left and right y-axis represents the $RMSE$ and $MAPE$ values respectively. The red dashed horizontal line represents the benchmark $RMSE$ value of 0.39. The combined legend in the upper right corner of each subplot shows the Goodness of Fit (GoF) for both $RMSE$ and MAPE.

Figure 9: The 50 shot CD-ViTO with ViT-B backbone predictions on the 1, 2, and 3 ID test dataset tile examples, respectively, from the left hand side to the right. The predicted bounding boxes in the images are the ones before non-maximum suppression and threshold. Black bounding boxes are the ground truth annotations, while the bounding boxes in the viridis color scale are the model predictions.

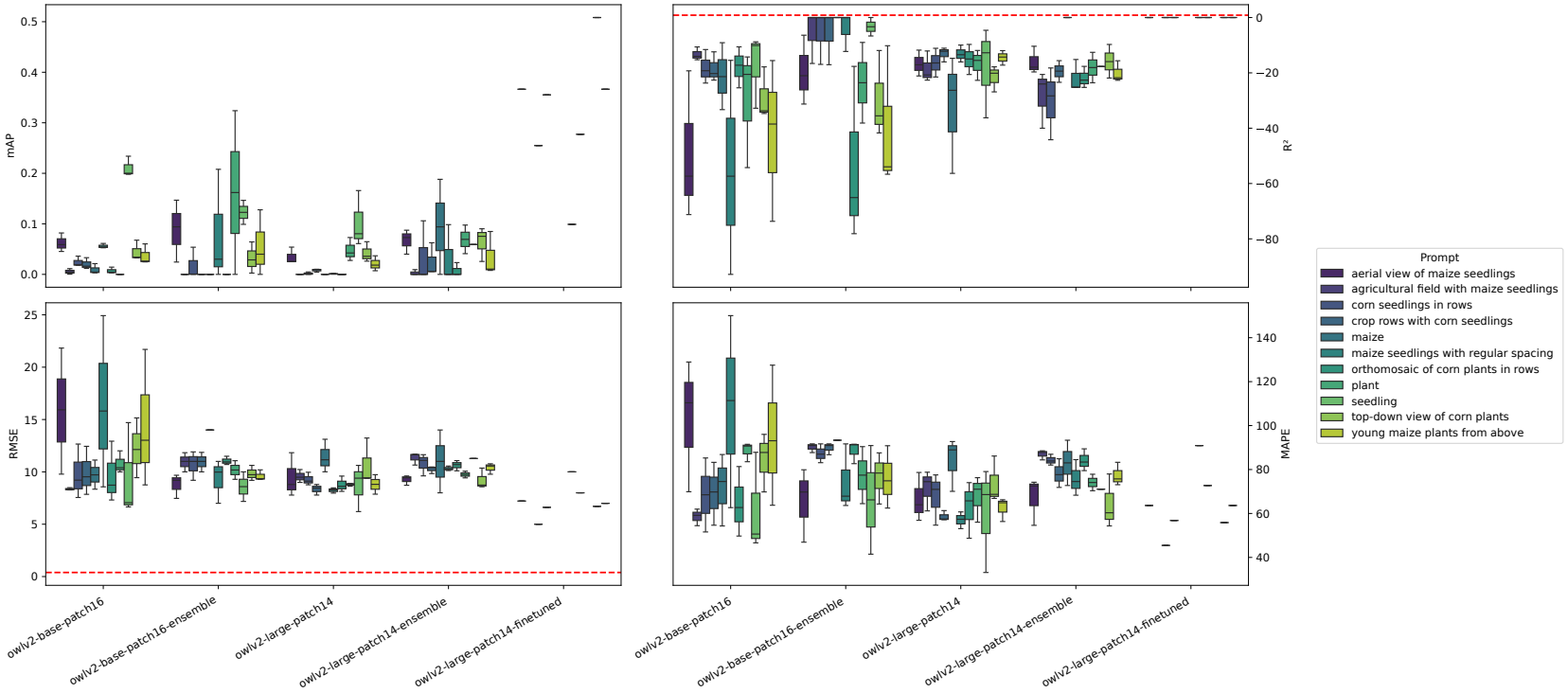

Figure 10: The figure shows the relationship between the OWLv2 model size, used prompt, and model performance. The x-axis represents the model settings and the model size. Colors represent the different prompts used. The four subplots show the mAP (upper left corner), R^2 (upper right corner), $RMSE$ (lower left corner), and MAPE (lower right corner) values. The red dashed horizontal line in the R^2 and the $RMSE$ subplots represents, respectively, the benchmarks of 0.85 and 0.39.

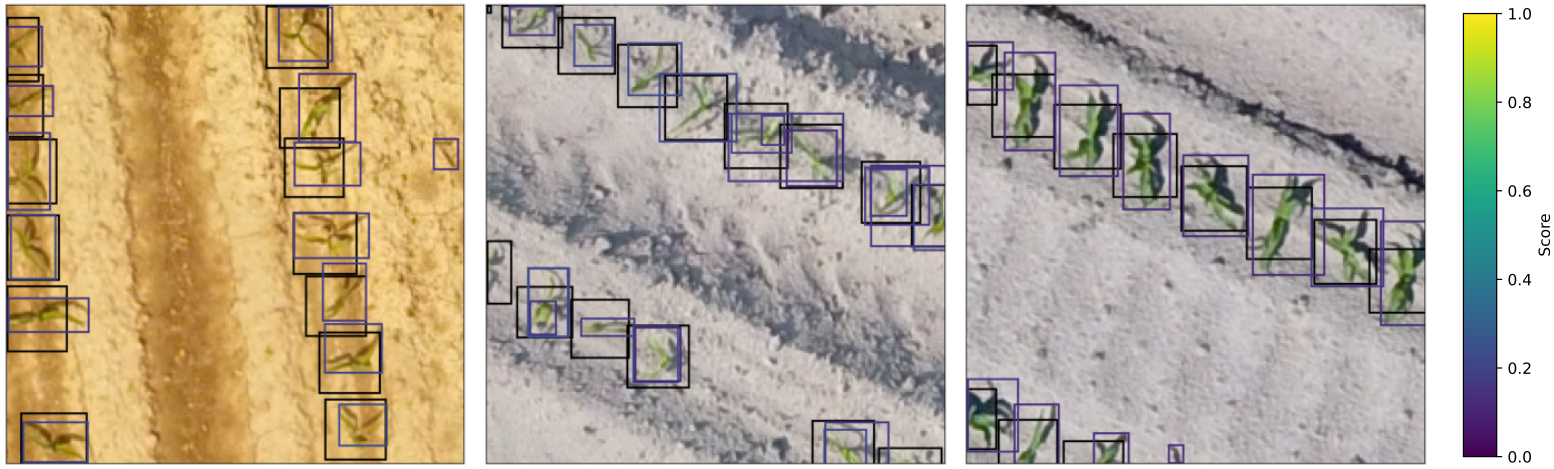

Figure 11: The best predictions with the OWLv2 model. The ID_1, ID_2, and ID_3 datasets, respectively, from the left hand side to the right. Prediction with owlv2-base-patch16-ensemble model of the ID_1 dataset, and with owlv2-base-patch16 model on the other two datasets. All the predictions are made with the prompt “seedling”. The predicted bounding boxes in the images are the ones before non-maximum suppression and threshold. Black bounding boxes are the ground truth annotations, while the bounding boxes in the viridis color scale are the model predictions.

2.1.4 Discussion

2.1.4.1 Dataset Source Impact on Object Detection Performance

Our experiments clearly demonstrate the critical importance of dataset source for successful arable crop seedling detection. None of the tested models, regardless of architecture or parameter amount, achieved the benchmark R^2 value of 0.85 when trained on Out-of-Distribution (OOD) datasets. Several inherent biases in our datasets likely influenced model performance. This aligns with previous findings by David et al. [16] and Andvaag et al. [47], who similarly reported significantly lower performance when using training samples from sources different from the inference dataset.

The domain gap challenge is particularly pronounced in agricultural applications, where environmental conditions, lighting, camera parameters, and plant growth stages vary substantially across datasets. The failure of OOD training highlights that visual features learned from one orthomosaic do not generalize well to others without significant adaptation. As the Goodness-of-Fit (GoF) of the models explicating the relationship between dataset size and performance was always below 0.2, one can argue that the interval of dataset size tested was too narrow to achieve a good fit or that other variables play an important role in determining model performance. Both cases are likely to be true, but also the maximum OOD dataset size that was tested (1168) was really small in respect to other studies that

use training datasets of tens of thousand of images to achieve such benchmarks [33]. This further highlights the importance of collecting in-domain training data, as the minimum OOD dataset size and quality to train an object detector to count arable crops seedling that generalizes to all the real-world cases is difficult to establish with a limited dataset.

Despite the poor performance of OOD dataset-trained models, some of them showed a low $MAPE$ value of less than 20%, not enough to consider the models for direct inferencing but rather as an annotation tool for the ID dataset.

2.1.4.2 Many-Shot Object Detection: Architecture and Dataset Requirements

Our results reveal important relationships between model architecture, count metrics, and minimum dataset requirements, consistent with findings from other agricultural computer vision studies. Within YOLO-family models, we observed that the lightweight YOLOv5n, YOLOv5s, and YOLOv8n achieved the benchmark with 130, 130, and 110 samples, respectively, well below the reported minimum by supposed few-shot studies with the same aim [43]. As already well-known from general computer vision literature [48, 52], increasing model complexity in CNN-based architectures corresponded to increased dataset size requirements.

Conversely, for transformer-mixed models like RT-DETR, we observed better performances with lower training dataset sizes, with RT-DETR

L achieving the benchmark with only 60 samples while the larger RT-DETR X required 100 samples. This superior performance of transformer-based models corroborates recent findings [34, 35], who demonstrated that attention mechanisms can achieve better performance with limited training data in agricultural contexts. The empirical models of dataset size versus performance showed comparable GoF between RT-DETR and YOLO-family models, except for YOLOv5n, which showed a particularly high GoF. This suggests that transformer-based models have the same predictability to reach the benchmark with the reported dataset size as the CNN-based models, except for YOLOv5n, which has a higher predictability to reach the benchmark given the same dataset size.

Overall, transformer-based models may require fewer samples to achieve the same performance as CNN-based models, potentially due to their ability to capture long-range dependencies and contextual information more effectively, as noted in a comprehensive review [32] and demonstrated in agricultural applications [33]. A visible side-effect of the adoption of transformer-mixed models is the higher computational cost of the training phase in terms of time and memory, which could be a limitation for some applications, consistent with computational trade-offs reported in [28] but not studied here.

This creates a practical tradeoff for practitioners: whether to use a simpler CNN-based model like YOLOv5n and invest in collecting more annotated images (approximately 130), or to allocate more computational resources for a transformer-mixed model like RT-DETR L that can achieve comparable performance with roughly half the

amount of labeled data (approximately 60 images). Similar trade-offs have been documented in agricultural deployment scenarios by [21] for early-season crop monitoring and [33] in their comprehensive review of agricultural object detection.

The predictability of model performance based on dataset size (as evidenced by GoF values exceeding 0.3 for successful models) provides practical guidance for practitioners. Our findings on dataset size scaling relationships (logarithmic patterns with GoF > 0.3) align with broader machine learning scaling laws established by [48, 50], while the agricultural-specific implications mirror domain adaptation challenges documented in [16, 47]. The relationship between dataset size and performance (modeled using logarithmic, arctangent, or algebraic root functions, depending on best fit) suggests diminishing returns beyond certain thresholds, consistent with theoretical frameworks proposed by [51], which can help inform efficient resource allocation for annotation efforts in precision agriculture applications.

2.1.4.3 Dataset Quality Trade-Offs

Our investigation into minimum dataset quality requirements revealed that models can tolerate some reduction in annotation quality while still maintaining benchmark performance achieved with the same training dataset size. YOLOv5n, YOLOv5s, and YOLOv8n achieved the benchmark with 85%, 90%, and 85% of the original dataset quality, while RT-DETR X required only 65%. Notably, RT-DETR L failed to maintain benchmark performance with any reduction in annotation quality, suggesting different sensitivity to annotation errors, consis-

tent with findings on annotation quality effects [49].

This difference in quality tolerance between RT-DETR L and RT-DETR X can be explained by considering their respective minimum dataset sizes. RT-DETR L was tested with quality reductions on its minimum benchmark-achieving dataset size of just 60 samples, while RT-DETR X was tested with 100 samples. With fewer training examples, RT-DETR L becomes more sensitive to the quality of each individual annotation, as each annotation represents a larger proportion of the total learning signal. In contrast, RT-DETR X, with its larger training dataset, can better compensate for quality reductions by leveraging redundancy across more examples, aligning with general principles of dataset robustness [48].

These findings provide valuable insights for practical applications, as they suggest that, in some cases, it may be more efficient to collect a larger quantity of moderate-quality annotations rather than focusing on perfect annotations for a smaller dataset. This also indicates potential for semi-automated annotation workflows, where machine assistance in annotation (which may introduce some errors) could be acceptable for many applications, supporting approaches documented in agricultural computer vision literature [33].

2.1.4.4 Few-Shot and Zero-Shot Approaches: Current Limitations

Despite recent advances in few-shot and zero-shot learning, our experiments reveal significant limitations in these approaches for pre-

cise maize seedling detection. The best CD-ViTO few-shot model achieved an $RMSE$ of 3.9 with 50 shots (using ViT-B backbone), substantially below the benchmark requirement of 0.39. Similarly, zero-shot models like OWLv2 performed poorly, regardless of prompt engineering efforts.

These results contrast with the promising performance reported for few-shot and zero-shot methods in general object detection benchmarks [41, 74]. Several factors may explain this gap: First, the domain-specific nature of aerial maize seedling imagery, where subtle textural differences and high intra-class variability are prevalent, severely challenges models pre-trained on general object detection datasets. As illustrated in the few-shot experiments (Figure 8), increasing the number of shots leads to nonlinear improvements in metrics such as R^2 and mAP (following an arctan-like trend), yet the error metric ($RMSE$) remain significantly above the benchmark. This saturation effect suggests that, even with more than usual maximum tested prototypes (50 instead of 30), the models struggle to capture the fine-grained visual cues necessary for precise seedling detection. Moreover, the zero-shot results (Figure 10) reveal a pronounced sensitivity to prompt phrasing, with all variants, including ensemble and fine-tuned versions of OWLv2, consistently failing to approach acceptable error rates. These observations imply that both the inherent complexity of the task and the limitations of current few-shot and zero-shot frameworks necessitate more domain-specific strategies, as suggested by recent work on agricultural computer vision challenges [14]. Addressing these challenges through domain-specific adaptations could help narrow the performance gap, potentially mak-

ing few-shot and zero-shot methods more competitive for arable crop seedling detection. Interestingly, the few-shot and the zero-shot models were able to detect all the seedlings without false positives in few cases. It would be interesting to investigate the possible ways to retain these images and use them to populate the training dataset for a many-shot model.

2.1.4.5 Handcrafted Methods in the Deep Learning Era

Despite the focus on deep learning approaches, our Handcrafted (HC) object detector demonstrated strong performance on the testing datasets (R^2 from 0.87 to 0.95). However, a significant limitation was the small proportion of tiles (1.8% to 7.8%) for which it could provide reliable annotations. This illustrates the classic trade-off of rule-based systems: high precision in constrained scenarios but limited generalizability, consistent with findings from other agricultural applications using handcrafted methods [22, 23].

These findings suggest that HC methods may still have value in a hybrid approach, where they provide high-quality annotations on a subset of data, which can then be used to bootstrap deep learning models. Such an approach could be particularly valuable for specialized agricultural applications where annotation resources are limited, aligning with hybrid approaches documented in crop monitoring studies [21].

This approach is highly adopted in industry, where the HC method is used to annotate the training dataset and the deep learning model is

used to predict the real-world cases, but it introduces a possible bias in the training dataset that could be a limitation for the model generalization. The main problem is that the HC1 method relies on color thresholding that filters the objects based on the color of the objects. That could be not the best way to annotate the training dataset for a deep learning model that could learn more complex features of the objects, but also the ones selected by the HC1 method.

2.1.4.6 Implications for Practical Applications

Our study has several practical implications for developing arable crop seedling detection systems. First, collecting in-domain training data is non-negotiable for achieving benchmark performance. Finding a way to automatically obtain the training dataset from the same distribution as the intended inference target is a key step in developing a robust object detector for arable crop seedling detection.

The logarithmic relationship between dataset size and performance suggests that initial annotation efforts should focus on reaching the minimum viable dataset size (60–130 images depending on architecture), after which additional annotations yield diminishing returns. This finding helps organizations optimize resource allocation for annotation efforts.

Our results also demonstrate that some reduction in annotation quality is acceptable, with models maintaining benchmark performance with 65–90% of the original quality. This suggests that semi-automated annotation workflows could be efficiently implemented for agricultural ap-

plications, potentially reducing the time and cost associated with manual annotation.

Current few-shot and zero-shot methods, while promising, are not yet viable replacements for traditional object detection approaches in seedling detection or counting tasks. However, they might still serve auxiliary roles in the annotation pipeline.

Hybrid approaches combining handcrafted methods with deep learning models could provide a practical solution for achieving benchmark performance. We observed that OOD many-shot, few-shot, and zero-shot models are occasionally able to produce annotations with sufficient quality for training ID many-shot models. A promising direction for future work would be to develop methods for automatically identifying and leveraging these high-quality annotations. Specifically, the HC2 component of our handcrafted approach could potentially be used to filter and validate annotations produced by these models, overcoming the color-thresholding bias introduced by HC1 while maintaining the agronomic knowledge encoded in HC2's row-pattern validation.

2.1.4.7 Future Work

In this study, we focused on the minimum dataset requirements for fine-tuning pre-trained models for the downstream task of counting arable crop seedlings through object detection. We did not explore the potential benefits of using domain-specific backbones. Future work could investigate whether or not dataset size requirements

could be further reduced by using backbones pre-trained on agricultural imagery, particularly aerial orthomosaics of crop fields. Such domain-specific pre-training might allow models to learn more relevant features for crop detection tasks, potentially reducing the amount of in-domain data needed for fine-tuning.

2.1.5 Conclusions

This study demonstrates that successful maize seedling detection requires in-domain training data, with out-of-distribution training requiring unreasonable dataset size to achieve benchmark performance across all tested models. We established minimum dataset requirements for several architectures, finding that lightweight YOLO models achieve benchmark performance with 110–130 samples, while certain transformer-mixed models like RT-DETR require as few as 60 samples. Models showed varying tolerance for reduced annotation quality, with some maintaining performance with only 65–90% of original annotation quality.

Despite advances in machine learning, neither few-shot nor zero-shot approaches currently meet precision requirements for arable crop seedling detection. Our handcrafted algorithm achieved excellent performance within its constraints, suggesting potential value in hybrid approaches combining rule-based methods with deep learning. These findings provide practical guidance for developing maize seedling detection systems, and possible ways to overcome the limitations of the current deep learning models for this application.

Author Contributions

Conceptualization, S.B. and E.B.-M.; methodology, S.B. and E.B.-M.; software, S.B.; validation, S.B.; formal analysis, S.B.; investigation, S.B.; resources, S.B. and E.B.-M.; data curation, S.B.; writing—original draft preparation, S.B.; writing—review and editing, E.B.-M.; visualization, S.B.; supervision, E.B.-M.; project administration, E.B.-M.; funding acquisition, S.B. and E.B.-M. All authors have read and agreed to the published version of the manuscript.

Funding

This research was conducted as part of a PhD program supported by SAGEA centro di saggio s.r.l.

The code for the handcrafted methods used in this study is available at `https://gist.github.com/SamueleBumbaca/4a227bbe7b78d6be3424899c16c60bb4` (accessed on 20 June 2025). The datasets created during this study (ID datasets) are available at the Zenodo repository `https://doi.org/10.5281/zenodo.15235602` (accessed on 20 June 2025). The other datasets used are available from their cited sources.

Conflicts of Interest

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpreta-

tion of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

R^2	Coefficient of determination
$RMSE$	Root Mean Squared Error
$MAPE$	Mean Absolute Percentage Error
mAP	Mean Average Precision
HC	Handcrafted
ID	In-Distribution
OOD	Out-of-Distribution
CNN	Convolutional Neural Network
ViT	Vision Transformer
YOLO	You Only Look Once
RT-DETR	Real-Time Detection Transformer
GoF	Goodness of Fit

no

2.1.6

2.1.6.1

Algorithm 1 H1

Require: $tiles$ ▷ Orthomosaic tiles

Require: col_range ▷ Color space thresholds

Require: $leaf\_area\_range$ ▷ Leaf area range in pixels

Ensure: $plants$ ▷ List of polygons

function connected_components($binary\_image$) [75]

    return $regions$

for $tile$ in $tiles$ **do**

    $mask \leftarrow \{p \in tile \mid color(p) \in col\_range\}$

    $regions \leftarrow connected\_components(mask)$

    $plants \leftarrow \{region \mid region \in regions \land region.area \in leaf\_area\_range\}$

return $plants$

Algorithm 2 H2

Require: $observations$ ▷ (List of centroids, RanSaC models)

Require: $intra - row\_dist$ ▷ Minimum distance between plants

Require: $inter - row\_dist$ ▷ Minimum distance between rows

Require: $mean\_slope$ ▷ Mean slope of the rows in respect meridian

Ensure: $objects$ ▷ List of centroids or polygons

function region_centroids($regions$) ▷ Get the centroids of the regions
    return $centroids$
function agglomerate_regions($regions, min\_dist$) ▷ Agglomerate regions
    $centroids \leftarrow \{region.centroid \mid region \in regions\}$
    $clusters \leftarrow$ HierarchicalClustering$(centroids, threshold = min_dist, metric =$ euclidean$)$
    $clust_cen \leftarrow \{$mean$(centroids_i) \mid$ for each cluster $i \in clusters\}$
    return $clust\_cen$
function extract_ransac_line($points, min\_dist$) [59]
    return $best\_inliers, best\_model$
function process_tiles($intra - row\_dist$)
    $observations \leftarrow \{\}$
    $plants \leftarrow$ HC1$(tiles)$
    for $tile$ in $tiles$ **do**
        $regions \leftarrow plants[tile]$
        $centroids \leftarrow$ region_centroids$(regions)$
        $clust\_cen \leftarrow$ agglomerate_regions$(regions, intra - row_dist)$
        $inlier\_points, model \leftarrow$ extract_ransac_line$(clust_cen, intra - row_dist)$
        $line\_length \leftarrow$ get_line_length$(model)$
        $expected\_number\_of\_plants \leftarrow \frac{line\_length}{intra\_row\_dist}$
        if $inlier\_points \equiv expected\_number\_of\_plants$ **then**
            $observations[tile] \leftarrow (clust\_cen, inlier\_points, model)$
    return $observations$

Algorithm 2 *Cont.*

```
function Filter_observations_by_slope(observations)
    filtered_observations ← {}
    for tile ∈ observations do
        slope ← observations[tile]['model']
        if model.slope ≈ mean_slope then
            filtered_observations[tile] ← observations[tile]
    return filtered_observations
function process_observations(observations, inter − row_dist, intra − row_dist)
    objects ← {}
    for tile ∈ observations do
        tile_centers ← observations[tile]['clust_cen']
        first_row_centers ← observations[tile]['inlier_points']
        first_row_model ← observations[tile]['model']
        centers ← {p | p ∈ tile_centers ∧ p ∉ first_row_centers}
        second_row_centers, second_row_model ← extract_ransac_line(centers, intra − row_dist)
        line_length ← get_line_length(second_row_model)
        expected_number_of_plants ← line_length / intra_row_dist
        if second_row_model.slope ≈ first_row_model.slope then
            if abs(second_row_model.intercept − first_row_model.intercept) ≈ inter − row_dist then
                if second_row_centers ≡ expected_number_of_plants then
                    objects[tile] ← (first_row_centers, second_row_centers)
    return objects
function main
    observations ← process_tiles(intra − row_dist)
    MEAN_SLOPE ← mean(observations['model'])
    observations ← Filter_observations_by_slope(observations, MEAN_SLOPE)
    objects ← process_observations(observations, inter − row_dist)
    return objects
```

2.1.6.2

The following are the list of prompts used for the zero-shot models:

- “maize”
- “seedling”
- “plant”
- “aerial view of maize seedlings”
- “corn seedlings in rows”
- “young maize plants from above”
- “crop rows with corn seedlings”
- “maize seedlings with regular spacing”
- “top-down view of corn plants”
- “agricultural field with maize seedlings”
- “orthomosaic of corn plants in rows”

References

Bibliography

[1] Blandino, M.; Testa, G.; Quaglini, L.; Reyneri, A. *Effetto Della DensitÃ Colturale e Dell'Applicazione di Fungicidi Sulla Produzione e la QualitÃ del Mais da Granella e da Trinciato*; ITA: Rome, Italy, 2016.

[2] Lu, D.; Ye, J.; Wang, Y.; Yu, Z. Plant Detection and Counting: Enhancing Precision Agriculture in UAV and General Scenes. *IEEE Access* **2023**, *11*, 116196–116205. [CrossRef]

[3] Saatkamp, A.; Cochrane, A.; Commander, L.; Guja, L.; Jimenez-Alfaro, B.; Larson, J.; Nicotra, A.; Poschlod, P.; Silveira, F.A.O.; Cross, A.; et al. A research agenda for seed-trait functional ecology. *New Phytol.* **2019**, *221*, 1764–1775. [CrossRef] [PubMed]

[4] De Petris, S.; Sarvia, F.; Gullino, M.; Tarantino, E.; Borgogno-Mondino, E. Sentinel-1 Polarimetry to Map Apple Orchard Damage after a Storm. *Remote Sens.* **2021**, *13*, 1030. [CrossRef]

[5] PP 1/333 (1) Adoption of Digital Technology for Data Gener-

ation for the Efficacy Evaluation of Plant Protection Products. *EPPO Bull.* **2025**, *55*, 14–19. [CrossRef]

[6] Zou, H.; Lu, H.; Li, Y.; Liu, L.; Cao, Z. Maize Tassels Detection: A Benchmark of the State of the Art. *Plant Methods* **2020**, *16*, 108. [CrossRef]

[7] Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2015**, arXiv:1405.0312. [CrossRef]

[8] Kraus, K. *Photogrammetry: Geometry from Images and Laser Scans*; De Gruyter: Berlin, Germany, 2011. [CrossRef]

[9] Pugh, N.A.; Thorp, K.R.; Gonzalez, E.M.; Elshikha, D.E.M.; Pauli, D. Comparison of image georeferencing strategies for agricultural applications of small unoccupied aircraft systems. *Plant Phenome J.* **2021**, *4*, e20026. [CrossRef]

[10] Dhonju, H.K.; Walsh, K.B.; Bhattarai, T. Web Mapping for Farm Management Information Systems: A Review and Australian Orchard Case Study. *Agronomy* **2023**, *13*, 2563. [CrossRef]

[11] Habib, A.; Han, Y.; Xiong, W.; He, F.; Zhang, Z.; Crawford, M. Automated Ortho-Rectification of UAV-Based Hyperspectral Data over an Agricultural Field Using Frame RGB Imagery. *Remote Sens.* **2016**, *8*, 796. [CrossRef]

[12] De Petris, S.; Sarvia, F.; Borgogno-Mondino, E. RPAS-based photogrammetry to support tree stability assessment: Longing for precision arboriculture. *Urban For. Urban Green.* **2020**, *55*, 126862. [CrossRef]

[13] Zhang, S.; Barrett, H.A.; Baros, S.V.; Neville, P.R.H.; Talasila, S.; Sinclair, L.L. Georeferencing Accuracy Assessment of Historical Aerial Photos Using a Custom-Built Online Georeferencing Tool. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 582. [CrossRef]

[14] Farjon, G.; Huijun, L.; Edan, Y. Deep-learning-based counting methods, datasets, and applications in agriculture: A review. *Precis. Agric.* **2023**, *24*, 1683–1711. [CrossRef]

[15] Meier, U.; Bleiholder, H.; Buhr, L.; Feller, C.; Hack, H.; Heß, M.; Lancashire, P.D.; Schnock, U.; Stauß, R.; van den Boom, T.; et al. The BBCH System to Coding the Phenological Growth Stages of Plants-History and Publications. *J. Für Kult.* **2009**, *61*, 41–52. [CrossRef]

[16] David, E.; Daubige, G.; Joudelat, F.; Burger, P.; Comar, A.; de Solan, B.; Baret, F. Plant Detection and Counting from High-Resolution RGB Images Acquired from UAVs: Comparison between Deep-Learning and Handcrafted Methods with Application to Maize, Sugar Beet, and Sunflower. *bioRxiv* **2021**. [CrossRef]

[17] Liu, W.; Zhou, J.; Wang, B.; Costa, M.; Kaeppler, S.M.; Zhang, Z. IntegrateNet: A Deep Learning Network for Maize Stand

Counting From UAV Imagery by Integrating Density and Local Count Maps. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6512605. [CrossRef]

[18] Maize_seeding Dataset > Overview. Available online: https://universe.roboflow.com/objectdetection-hytat/maize_seeding (accessed on 20 June 2025).

[19] Maize-Seedling-Detection Dataset > Overview. Available online: https://universe.roboflow.com/fyxdds-icloud-com/maize-seedling-detection (accessed on 20 June 2025).

[20] FAO. *Agricultural Production Statistics 2010–2023*; Volume Analytical Briefs; FAOSTAT: Rome, Italy, 2024.

[21] Torres-SÃ¡nchez, J.; Mesas-Carrascosa, F.J.; JimÃ©nez-Brenes, F.M.; de Castro, A.I.; LÃ³pez-Granados, F. Early Detection of Broad-Leaved and Grass Weeds in Wide Row Crops Using Artificial Neural Networks and UAV Imagery. *Agronomy* **2021**, *11*, 749. [CrossRef]

[22] Zhang, Z.; Cao, R.; Peng, C.; Liu, R.; Sun, Y.; Zhang, M.; Li, H. Cut-edge detection method for rice harvesting based on machine vision. *Agronomy* **2020**, *10*, 590. [CrossRef]

[23] García-Martínez, H.; Flores-Magdaleno, H.; Khalil-Gardezi, A.; Ascencio-Hernández, R.; Tijerina-Chávez, L.; Vázquez-Peña, M.A.; Mancilla-Villa, O.R. Digital Count of Corn Plants Using Images Taken by Unmanned Aerial Vehicles and Cross Correlation of Templates. *Agronomy* **2020**, *10*, 469. [CrossRef]

[24] LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]

[25] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks|IEEE Journals & Magazine|IEEE Xplore. Available online: https://ieeexplore-ieee-org.bibliopass.unito.it/document/7485869 (accessed on 20 June 2025).

[26] You Only Look Once: Unified, Real-Time Object Detection|IEEE Conference Publication|IEEE Xplore. Available online: https://ieeexplore-ieee-org.bibliopass.unito.it/document/7780460 (accessed on 20 June 2025).

[27] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; NIPS'17, pp. 6000–6010.

[28] Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872. [CrossRef]

[29] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]

[30] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2014**, *115*, 211â€“252. [CrossRef]

[31] Zong, Z.; Song, G.; Liu, Y. DETRs with Collaborative Hybrid Assignments Training. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 6725–6735. [CrossRef]

[32] Khan, A.; Rauf, Z.; Sohail, A.; Khan, A.R.; Asif, H.; Asif, A.; Farooq, U. A Survey of the Vision Transformers and Their CNN-transformer Based Variants. *Artif. Intell. Rev.* **2023**, *56*, 2917–2970. [CrossRef]

[33] Badgujar, C.M.; Poulose, A.; Gan, H. Agricultural Object Detection with You Only Look Once (YOLO) Algorithm: A Bibliometric and Systematic Literature Review. *Comput. Electron. Agric.* **2024**, *223*, 109090. [CrossRef]

[34] Rekavandi, A.M.; Rashidi, S.; Boussaid, F.; Hoefs, S.; Akbas, E.; Bennamoun, M. Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art. *arXiv* **2023**, arXiv:2309.04902. [CrossRef]

[35] Li, Y.; Miao, N.; Ma, L.; Shuang, F.; Huang, X. Transformer for Object Detection: Review and Benchmark. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107021. [CrossRef]

[36] Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. *arXiv* **2024**, arXiv:2304.08069. [CrossRef]

[37] Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]

[38] Li, Z.; Zhou, F.; Chen, F.; Li, H. Meta-SGD: Learning to Learn Quickly for Few-Shot Learning. *arXiv* **2017**, arXiv:1707.09835. [CrossRef]

[39] Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; Divakaran, A. Zero-Shot Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 384–400.

[40] Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-Shot Object Detection via Feature Reweighting. *arXiv* **2019**, arXiv:1812.01866. [CrossRef]

[41] Minderer, M.; Gritsenko, A.; Houlsby, N. Scaling Open-Vocabulary Object Detection. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 72983–73007.

[42] Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In Proceedings of the Computer Vision—ECCV 2024, Milan, Italy, 29 September–4 October 2024; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer: Cham, Switzerland, 2025; pp. 38–55. [CrossRef]

[43] Karami, A.; Crawford, M.; Delp, E.J. Automatic Plant Counting and Location Based on a Few-Shot Learning Technique. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5872–5886. [CrossRef]

[44] Wang, D.; Parthasarathy, R.; Pan, X. Advancing Image Recognition: Towards Lightweight Few-shot Learning Model for Maize Seedling Detection. In Proceedings of the 2024

International Conference on Smart City and Information System, Kuala Lumpur, Malaysia, 17–19 May 2024; pp. 635–639. [CrossRef]

[45] Barreto, A.; Lottes, P.; Ispizua Yamati, F.R.; Baumgarten, S.; Wolf, N.A.; Stachniss, C.; Mahlein, A.K.; Paulus, S. Automatic UAV-based Counting of Seedlings in Sugar-Beet Field and Extension to Maize and Strawberry. *Comput. Electron. Agric.* **2021**, *191*, 106493. [CrossRef]

[46] Kitano, B.T.; Mendes, C.C.T.; Geus, A.R.; Oliveira, H.C.; Souza, J.R. Corn Plant Counting Using Deep Learning and UAV Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1–5. [CrossRef]

[47] Andvaag, E.; Krys, K.; Shirtliffe, S.J.; Stavness, I. Counting Canola: Toward Generalizable Aerial Plant Detection Models. *Plant Phenomics* **2024**, *6*, 0268. [CrossRef]

[48] Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *arXiv* **2017**, arXiv:1707.02968. [CrossRef]

[49] Alhazmi, K.; Alsumari, W.; Seppo, I.; Podkuiko, L.; Simon, M. Effects of Annotation Quality on Model Performance. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 063–067. [CrossRef]

[50] Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M.M.A.; Yang, Y.; Zhou, Y. Deep Learning Scaling Is Predictable, Empirically. *arXiv* **2017**, arXiv:1712.00409. [CrossRef]

[51] Mahmood, R.; Lucas, J.; Acuna, D.; Li, D.; Philion, J.; Alvarez, J.M.; Yu, Z.; Fidler, S.; Law, M.T. How Much More Data Do I Need? Estimating Requirements for Downstream Tasks. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 275–284. [CrossRef]

[52] Nguyen, N.D.; Do, T.; Ngo, T.D.; Le, D.D.; Valenti, C.F. An Evaluation of Deep Learning Methods for Small Object Detection. *JECE* **2020**, *2020*, 8856387. [CrossRef]

[53] Du, X.; Lin, T.Y.; Jin, P.; Ghiasi, G.; Tan, M.; Cui, Y.; Le, Q.V.; Song, X. SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11592–11601.

[54] Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]

[55] Liu, S.; Yin, D.; Feng, H.; Li, Z.; Xu, X.; Shi, L.; Jin, X. Estimating Maize Seedling Number with UAV RGB Images and

Advanced Image Processing Methods. *Precis. Agric.* **2022**, *23*, 1604–1632. [CrossRef]

[56] Velumani, K.; Lopez-Lozano, R.; Madec, S.; Guo, W.; Gillet, J.; Comar, A.; Baret, F. Estimates of Maize Plant Density from UAV RGB Images Using Faster-RCNN Detection Model: Impact of the Spatial Resolution. *Plant Phenomics* **2021**, *2021*, 9824843. [CrossRef]

[57] Bumbaca, S. The Original Dataset for the Paper “On the minimum dataset requirements for fine-tuning an object detector for arable crop plant counting: A case study on maize seedlings”. *Zenodo* **2025**. [CrossRef]

[58] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2012; Volume 25.

[59] Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In *Readings in Computer Vision*; Fischler, M.A., Firschein, O., Eds.; Morgan Kaufmann: San Francisco, CA, USA, 1987; pp. 726–740. [CrossRef]

[60] Terven, J.; CÃ³rdova-Esparza, D.M.; Romero-GonzÃ¡lez, J.A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]

[61] Jocher, G.; Qiu, J.; Chaurasia, A. GitHub Ultralytics YOLO. 2023. Available online: `https://github.com/ultralytics/ultralytics` (accessed on 16 April 2025).

[62] Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2024**, arXiv:2304.07193. [CrossRef]

[63] Fu, Y.; Wang, Y.; Pan, Y.; Huai, L.; Qiu, X.; Shangguan, Z.; Liu, T.; Fu, Y.; Gool, L.V.; Jiang, X. Cross-Domain Few-Shot Object Detection via Enhanced Open-Set Object Detector. *arXiv* **2024**, arXiv:2402.03094. [CrossRef]

[64] Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.

[65] Zhu, H.; Qin, S.; Su, M.; Lin, C.; Li, A.; Gao, J. Harnessing Large Vision and Language Models in Agriculture: A Review. *arXiv* **2024**, arXiv:2407.19679. [CrossRef]

[66] Zhou, Y.; Yan, H.; Ding, K.; Cai, T.; Zhang, Y. Few-Shot Image Classification of Crop Diseases Based on Visionâ€“Language Models. *Sensors* **2024**, *24*, 6109. [CrossRef]

[67] Chen, H.; Huang, W.; Ni, Y.; Yun, S.; Liu, Y.; Wen, F.; Velasquez, A.; Latapie, H.; Imani, M. TaskCLIP: Extend Large Vision-Language Model for Task Oriented Object Detection. *arXiv* **2024**, arXiv:2403.08108. [CrossRef]

[68] Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef]

[69] Draper, N.R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1998; Volume 326.

[70] Armstrong, J.; Collopy, F. Error measures for generalizing about forecasting methods: Empirical comparisons. *Int. J. Forecast.* **1992**, *8*, 69–80. [CrossRef]

[71] Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]

[72] Vianna, L.S.; GonÃ§alves, A.L.; Souza, J.A. Analysis of learning curves in predictive modeling using exponential curve fitting with an asymptotic approach. *PLoS ONE* **2024**, *19*, e0299811. [CrossRef]

[73] Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 966–970. [CrossRef]

[74] Xu, G.; Hao, Z.; Luo, Y.; Hu, H.; An, J.; Mao, S. DeViT: Decomposing Vision Transformers for Collaborative Inference in Edge Devices. *arXiv* **2023**, arXiv:2309.05015. [CrossRef]

[75] Wu, K.; Otoo, E.; Shoshani, A. Optimizing Connected Component Labeling Algorithms. In Proceedings of the Medical Imaging 2005: Image Processing, San Diego, CA, USA, 12–17 February 2005.

3.2 Ordinal Variables: Phytotoxicity Scoring Automation

Note: The following section is based on published work:

Bumbaca, S.; Borgogno-Mondino, E.C. Supporting Screening of New Plant Protection Products through a Multispectral Photogrammetric Approach Integrated with AI. Agronomy 2024, 14, 306. DOI: 10.3390/agronomy14020306

This study evaluates the effectiveness of digital techniques for reproducing ordinal scale assessments traditionally performed manually, specifically phytotoxicity scoring. The research explores how machine learning combined with photogrammetric and spectral imaging data can automate subjective visual assessments while maintaining EPPO compliance standards.

Abstract

This work was aimed at developing a prototype system based on multispectral digital photogrammetry to support tests required by international regulations for new Plant Protection Products (PPPs). In particular, the goal was to provide a system addressing the challenges of a new PPP evaluation with a higher degree of objectivity with respect to the current one, which relies on expert evaluations. The system uses Digital Photogrammetry, which is applied to multispectral acquisitions and Artificial Intelligence (AI). The goal of this paper is also to simplify the present screening process, moving it towards more objective and quantitative scores about phytotoxicity. The implementation of an opportunely trained AI model for phytotoxicity prediction aims to convert ordinary human visual observations, which are presently provided with a discrete scale (forbidding a variance analysis), into a continuous variable. The technical design addresses the need for a reduced dataset for training the AI model and relating discrete observations, as usually performed, to some proxy variables derived from the photogrammetric multispectral 3D model. To achieve this task, an appropriate photogrammetric multispectral system was designed. The system operates in multi-nadiral-view mode over a bench within a greenhouse exploiting an active system for lighting providing uniform and diffuse illumination. The whole system is intended to reduce the environmental variability of acquisitions tending to a standard situation. The methodology combines

advanced image processing, image radiometric calibration, and machine learning techniques to predict the General Phytotoxicity percentage index (PHYGEN), a crucial measure of phytotoxicity. Results show that the system can generate reliable estimates of PHYGEN, compliant with existing accuracy standards (even from previous PPPs symptom severity models), using limited training datasets. The proposed solution addressing this challenge is the adoption of the Logistic Function with LASSO model regularization that has been shown to overcome the limitations of a small sample size (typical of new PPP trials). Additionally, it provides the estimate of a numerical continuous index (a percentage), which makes it possible to tackle the objectivity problem related to human visual evaluation that is presently based on an ordinal discrete scale. In our opinion, the proposed prototype system could have significant potential in improving the screening process for new PPPs. In fact, it works specifically for new PPPs screening and, despite this, it has an accuracy consistent with the one ordinarily accepted for human visual approaches. Additionally, it provides a higher degree of objectivity and repeatability.

2.2.1 Introduction

Researchers in the field of Plant Protection Products (PPPs) need to bridge the gap between evaluations from traditional human-based approaches and those enabled by Artificial Intelligence (AI) [22]. Specifically, new PPPs undergo a rigorous safety screening before market entry. PPP developers must meticulously formulate and dose these PPPs to avoid harmful phytotoxic effects on crops, thus maintaining selectivity [21]. Traditionally, experimenters assess the severity of phytotoxicity through visual observations. The reliability of these assessments depends on low variability among experimenters' observations and proper rating scales [17]. In Europe, technicians are required to operate according to Good Experimental Practice (GEP), which is based on international laws [2]. GEP is a set of standards that ensures objectivity and precision in scientific experiments. The World Trade Organization Agreement on Sanitary and Phytosanitary Measures [9] designates the International Plant Protection Convention (IPPC) as the authority for plant health standards [45]. The European Union falls under the European and Mediterranean Plant Protection Organization (EPPO) within IPPC. EPPO is responsible for setting phytosanitary and PPP standards. EPPO standards address crop selectivity [5] by providing evaluation methods involving both discrete and continuous values. However, experimenters often prefer using quantitative ordinal discrete scales due to their practicality [18]. As observed by Chiang et al. [17], percentage scales with intervals of 10% can reduce rater uncertainty. That is because 10% is commonly accepted as inter-rater error. This can

potentially lead to inconsistencies with theoretical assumptions in variance analysis [48, 8]. Nevertheless, the selectivity of PPPs is inherently a continuous variable, assumed to be inversely proportional to the percentage of phytotoxicity symptoms and their intensity. According to EPPO, phytotoxicity symptoms include (i) modifications in the development cycle, (ii) thinning, (iii) modifications in color, (iv) necrosis, (v) deformation, and (vi) effects on quantity and quality of the yield [5]. General Phytotoxicity (PHYGEN) is an aggregate indicator that summarizes the above symptoms by defining the percentage of damage to a plant compared to a perfectly healthy reference plant [44].

Imaging sensors have already been demonstrated to improve precision and objectivity in the detection of pathological symptoms [18, 39]. Some spectral properties of plants, as recorded through multispectral sensors [38] are recognized as indicators of photosynthetic efficiency [25, 16]. Various methods, including multi-view approaches [46, 35, 53], can be used to create 3D models of plants [39]. Spectral and geometric features of plants can be used to virtually reproduce the plant appearance, as observed by an experimenter during assessment. When working with three-dimensional and multispectral data, a summary is necessary to obtain an accurate estimate of PHYGEN, like a direct human-based evaluation approach. Machine learning (ML) models from artificial intelligence (AI) can synthesize vast amounts of digital information in a robust and reasonable manner when guided by expert (low variation) experimenter annotations [38]. Open platforms offer large labeled training datasets, allowing users to customize ML algorithms to their re-

quirements [32, 29] Convolutional Neural Networks (CNNs) were found to be the most accurate method for symptom classification [50, 41] while working with image-based data. CNNs were shown to be capable of rating EPPO symptoms, specifically “modifications in color”, at both leaf and canopy levels [26]. Gómez-Zamanillo et al. [28] proposed a method for assessing PHYGEN by classifying the most common symptoms. Their study demonstrated the effectiveness of CNNs as feature extractors for predicting PHYGEN rates or similar measures. The study utilized CNN to identify and classify color-related phytotoxicity symptoms from RGB images. Severity estimates were determined by assigning arbitrary weights to the detected symptoms. Rather, they relied on expert experimenters to quantify weights without optimizing scores. Currently, no CNN-based model has been proposed to generate a reasonable estimate of PHYGEN based on a comprehensive analysis of all symptoms. Weight optimization is highly appreciated as it is expected to enhance the accuracy of estimates and provide insights into the significance of each symptom in the toxicological mechanism of PPPs. Further challenges associated with the deployment of CNNs for plant disease detection and scoring are reported in Barbedo et al. [13, 12]. In particular, these include (i) sensitivity of deductions to environmental and sensor-related issues, (ii) capability of generalization of the model, and (iii) training dataset quality. It is important to note that the quality of the training dataset is highly significant as it must be properly calibrated for the specific type of PPP being tested. Therefore, pre-trained networks relying on training datasets generated for different symptoms from different PPPs should not be used to test

new PPPs. It is worth noting that, in order for CNN training to be robust and accurate enough, it requires huge training datasets consisting of thousands of images. Table 1 shows some of the methods proposed in the literature for the estimation of PHYGEN, enhancing their suitability for new PPPs PHYGEN prediction.

Table 1: Related works.

Paper	Method	Accuracy[1]	Suitability[2]
Human raters	Depending on the rater, the recommended maximum error is 10% [17]	Traditional method	-
Ali et al. [10] no AI involved, and no monitorable stability	Image processing	Not reported	Involving only biomass estimation,
Chu et al. [19]	Shallow CNN	80%	Destructive and only spectral signature involved
Ghosal et al. [26]	CNN	From 50% to 90% depending on rater	Not phytotoxicity-specific, destructive
Gómez-Zamanillo et al. [28]	CNN	93.26%	Not suitable for new PPPs because of the amount of training data required

[1] It indicates the accuracy of phytotoxicity severity with respect to human raters.

[2] For new PPPs PHYGEN screening.

Typical trials for new PPPs usually involve only a few hundred plants. This may not provide a sufficient dataset for robust training, testing, and deployment of a new CNN. It is noteworthy that CNNs maintain their efficacy when symptoms of phytotoxicity are well-documented and recognized within the training dataset. This specificity is a true challenge in ML optimization for the newer PPP-related trials since the explored symptomatology may not be cataloged. This work emphasizes that symptoms of phytotoxicity resulting from new PPPs can be unique due to their novelty, making them unpredictable. Therefore, screening trials are necessary. The proposed method involves a PHYGEN evaluation via a CV ML system for new PPPs operating in a greenhouse environment that overcome such limitations. The system is specifically designed to address three key challenges in adopting AI, and specifically CV ML for new PPPs screening: small amount of training data, stability, and accuracy. Moreover, the model prediction suitability for ANOVA testing is also discussed. The presented method requires only a small training sample with respect to CNN algorithms because it relies on a single linear regression and a logistic function. It takes a small training sample from the available study population, effectively addressing issues of under-representation of training datasets [13], which is typical when testing new PPP phytotoxicity.

The system was found to reduce the impact of environmental and sensor-related factors on plant symptom detection, increasing the stability of plant pictures and measures. This is achieved through proper platform calibration techniques and a multi-view image capture approach that allows for the monitoring of errors of the geo-

metrical and radiometric measures used to train and test the model. Model stability was tested using cross-validation. The results confirmed the robustness of the method regardless of the sample adopted. The accuracy of the model's prediction was compared to the precision of human raters as described in the literature (10%) [17] and to the state-of-the-art (SOTA) model for PHYGEN of non-new PPPs (6.74%) [28]. It was not possible to find a direct comparison of a model predicting PHYGEN for new PPPs by CV ML in the literature. Therefore, the accuracy must be considered satisfactory if it is higher than the precision of human raters, and it is expected to be lower than that of CNN models with a greater amount of training data. The methodology also addresses the challenge of adopting discrete quantitative scales in the ML training step. It has been shown to improve the prediction of PHYGEN as a continuous scale variable, starting from quantitative ordinal discrete values, such as those obtained from ordinary approaches. Furthermore, as the PHYGEN estimates are now on a continuous scale, the ANOVA test can be more appropriately utilized, resolving the cumbersome lack of adaptation to the statistical theory that is often observed in the field of PPP screening.

2.2.2 Materials and Methods

2.2.2.1 Hardware Platform

A platform was developed and integrated into a greenhouse structure for multispectral photogrammetric data acquisition. The integration was achieved using a framework consisting of two 10-m-long aluminum extruded profiles affixed to the roof and walls of the greenhouse. To enable the sensing system to move along the Y-axis, two parallel linear rail guides were mounted on these profiles. In addition, a 6-m-long aluminum support was installed perpendicular to the initial rails. This support incorporates a linear guide rail, which enables camera movement along the X-axis. Adjustments along the Z-axis were made possible by altering the brackets on the Y-axis rails. The proximity of the sensing system to the bench, where the pots were situated, was adjustable within a range of 1.1 to 1.5 m. The camera's position along the 6-m rail could be adjusted using fixing brackets, as shown in fig. 1.

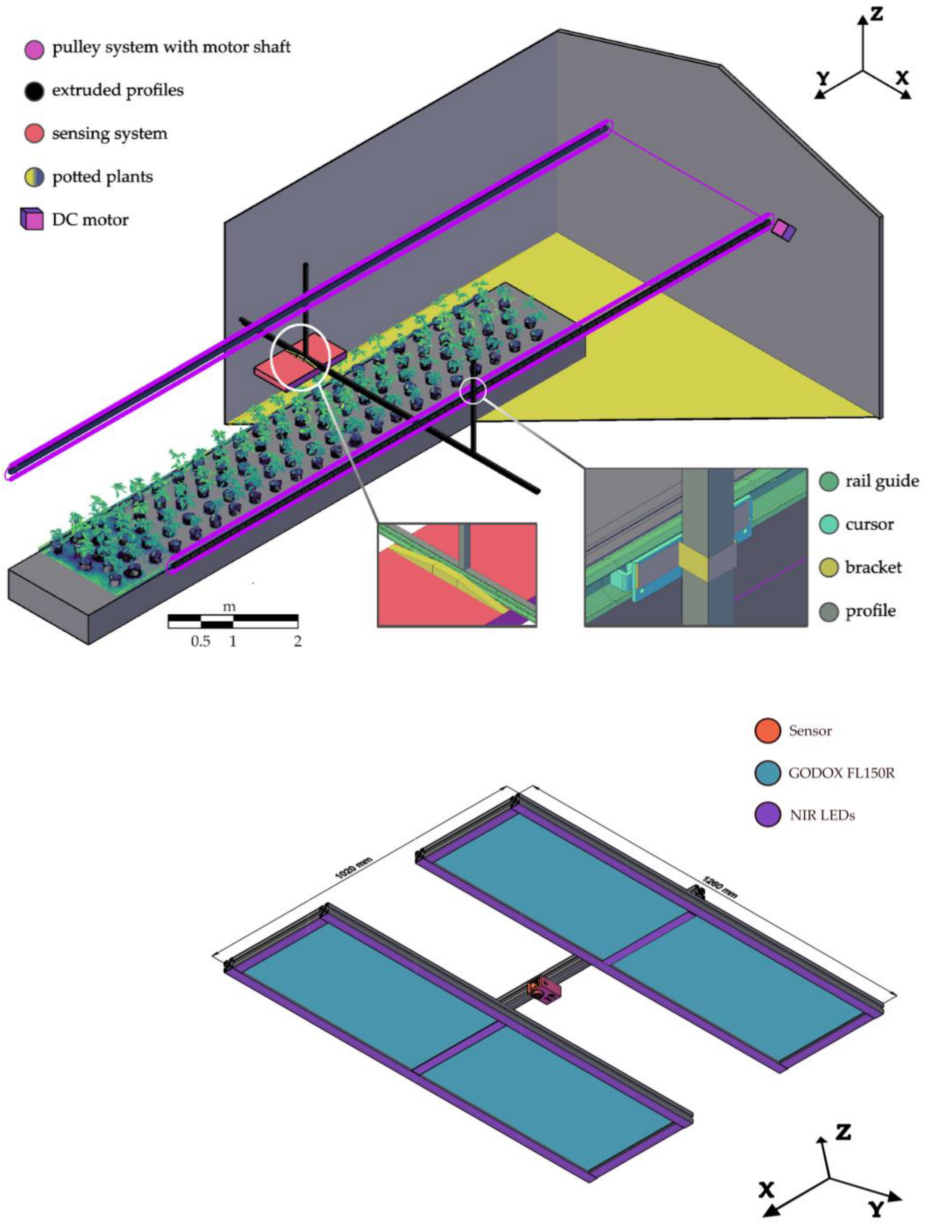

Figure 1: Platform and sensing system (top) and only the sensing system (bottom).

Camera movement along the Y-axis in the greenhouse was controlled by a DC motor that operates through a pulley system. This

system works similarly to a bridge crane that moves the imaging compound automatically with a speed of about 0.08 m/s along the Y-axis. The motion was manually activated and stopped. The whole moving platform was made of (i) one MAPIR Survey3W (PeauProductions, San Diego, CA, USA) camera (S3) multispectral camera, (ii) two Light-Emitting Diode (LED) panels (GODOX FL150R) (Godox, Shenzhen, Guangdong, China) each measuring 1.2 × 0.3 m, and (iii) a 6-m LED strip emitting with a peak at 850 nm that encircles the GODOX FL150R panels to ensure that adequate Near-Infrared (NIR) radiation reaches the plants. Panels (the entire imaging system) ran parallelly to the bench hosting the pots to be imaged to ensure uniform illumination. Furthermore, shading curtains were installed on the walls and ceiling of the greenhouse to reduce exterior light contribution during data collection. A preliminary test was conducted to ensure the consistency of the spectrum provided by LEDs through its comparison with the reflectance spectrum acquired by an RS-5400 Spectroradiometer (Spectral Evolution, Haverhill, MA, USA). The acquisition was performed using calibrated panels of the RS-5400 instrument fig. 2 in lighting conditions replicating the operational environment.

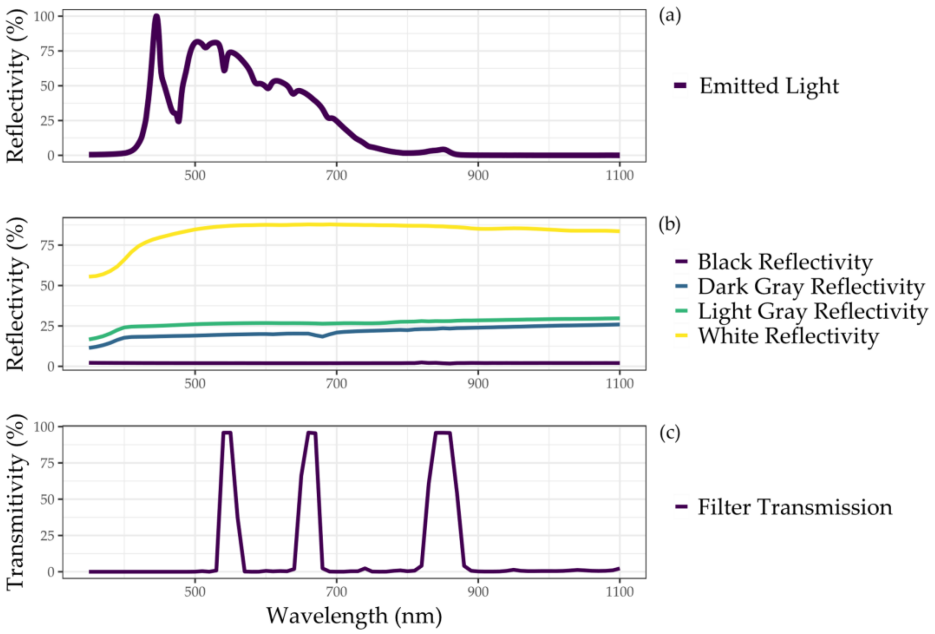

Figure 2: (a) Spectral signature of the reference panel, lighted with the tested LEDs and measured using the RS-5400 Spectro-radiometer. (b) Reflectivity of the MAPIR calibration panels corresponding to the different grayscale levels (yellow, light-green, blue, and violet colors in the graph) provided by the factory. The dark green line shows the filter sensitivity of MAPIR for the different bands. (c) Transmissivity of the S3 camera filter.

The S3 camera was used for image capture, as detailed in table 2. A white balance setting was employed during acquisition to increase the intensity of the Red and NIR bands, resulting in a reduction of green band sensitivity.

Table 2: S3 and system integration specifics.

Parameter	Value
Focal Length	3.37 mm
Aperture	f/2.8 (fixed)
Lens Distortion	<1%
Focal Length	Fixed
Hyper-focal Distance	81.5 cm
Sensor Size	3000 × 4000 pixels
Pixel Physical Size	1.55 µm
Bands	Green, Red, and NIR (Figure 2c)
Camera Shift (Y-axis)	20.3 cm per shot
Frames per second	1/3
Horizontal Footprint[1]	202–276 cm
Vertical Footprint[1]	152–207 cm

[1] at a 1.1–1.5 m distance.

2.2.2.2 Experimental Design

An experiment was conducted to assess the reliability of the system and the processing workflow with respect to EPPO standards. The selectivity of a herbicide with an unknown mode of action was tested in a controlled environment greenhouse following EPPO standards [5, 4, 6, 7]. This allowed uniform growing conditions to be maintained throughout the greenhouse. Forty-four pots, each 40 × 30

cm, were sown with oilseed rape (OSR) and treated with the experimental product before emergence. The treatments were applied using an automatic spray chamber. To ensure a balanced set of PHYGEN, different concentrations of the herbicide, including a control group, were used to cover a range of phytotoxicity intensities. Visual and digital evaluations were carried out simultaneously. The PHYGEN assessment values were recorded as Day After Treatment (DAA) in table 3.

Table 3: PHYGEN observations.

DAA[1]	**0%**	**13%**	**38%**	**63%**	**88%**
3	11	9	8	7	9
7	5	4	15	10	10
14	15	14	9	6	0
TOT	31	27	32	23	19

[1] Days After Application.

Only five discrete PHYGEN values were retained for scoring: 0%, 13%, 38%, 63%, and 88%. This emphasizes the nature of the data generated by the visual assessment and the extreme use of the discrete quantitative scale. It is important to note that all five values were assigned during the three assessments, except on the last day, when the highest value (88%) was not observed. This resulted in an imperfectly balanced distribution of PHYGEN over time. The interval between consecutive discrete values was 25%, except for the interval between 0% and 13%. The 0% value may be unreliable for treated pots due to the inevitable effect of herbicides, even for re-

sistant crops. The true value in the range between 0% and 13% is difficult to detect, even for expert experimenters upon visual inspection, and is usually interpreted as having no effect on the harvest. Despite this, 0% values were always recorded, as assessed by the experimenters.

2.2.2.3 Data Processing

The workflow starts with planning the image acquisition of the experimental plants. Then, the images are used to retrieve the multispectral 3D reconstruction of the plants. The parameters of the observed plants are extracted by the 3D model. Finally, the ML model is trained on the extracted parameters and validated. The workflow is summarized in fig. 3.

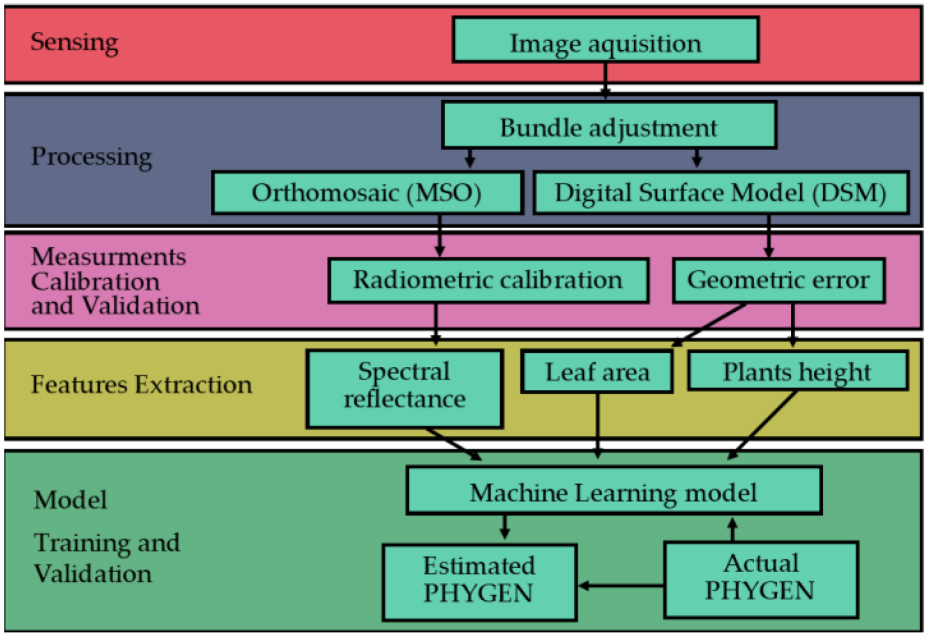

Figure 3: General workflow of the suggested method.

2.2.2.4 Planning the Acquisition

The camera movements were planned to capture stereoscopic images using a local Euclidean coordinate system, hereinafter called Coordinate Reference System (CRS), having the origin located at the lower left corner of the bench hosting plants. Image block bundle adjustment was intended to refine both position and attitude image Exterior Orientation (EO) parameters, using nominal coordinates of the focal point position and a nadiral orientation as an initial solution during the adjustment. Nominal values for image focal point position (X0, Y0, Z0) were determined assuming (i) X0 as the horizontal distance between adjacent strips; (ii) Y0 was computed by considering the speed of the camera shifts along the bars, and (iii) Z0 was set to a fixed value, which is discussed in the next paragraph. The camera was positioned with the longer side (4000 pixels) aligned across the track. The nominal Z0 value was determined based on two conditions. First, the resulting image footprint must be consistent with the expected target size of plants. Second, targets should be visible at the smallest distance longer than the hyper-focal distance (0.815 m for S3). This condition ensures the maximum obtainable resolution, known as Ground Sampling Distance (GSD), which maximizes the efficiency and quality of tie point recognition. It is important to note that GSD is proportional to the physical pixel size according to eq. (1).

$$GSD = \delta \frac{H}{f} \tag{1}$$

where H is the camera-to-target distance, f is the camera focal length, and δ is the physical pixel size. As the height of the assessed plants can vary greatly during the same acquisition, H can range from 0.815 to 1.500, resulting in a ground sample distance (GSD) that varies between 0.37 and 0.69 mm·pixel^{-1}.

When planning an acquisition, it is important to ensure that the coarser GSD (which depends on H) is smaller than the smallest feature that needs to be recognized. Tie point recognition depends on both the forward and side overlap among images. The forward overlap is determined by the baseline (B), which is the distance between consecutive focal points along the same strip. On the other hand, the side overlap is determined by the distance between two adjacent strips. The platform is designed to operate with a strip distance equal to the baseline (0.2 m), resulting in 95In digital photogrammetry, it is widely acknowledged that the Z coordinate of target points is the most critical to estimate accurately. Its precision can be evaluated using eq. (2) [20, 14, 34].

$$\sigma_z = \frac{H^2}{Bf}\sigma_x \tag{2}$$

where H is the camera-target distance, σ_x is the precision of parallax measures in the image domain (assumed to be half the physical pixel size, i.e., 1.685 µm for S3), B is the baseline, f is the sensor focal length, and σ_z is the estimated precision of the Z coordinate of the target point. The graphs in fig. 4 relate the theoretical (expected) σ_z with the baseline B while varying the camera-to-target distance at three reference values. The B interval was considered to be within

the minimum and maximum overlap.

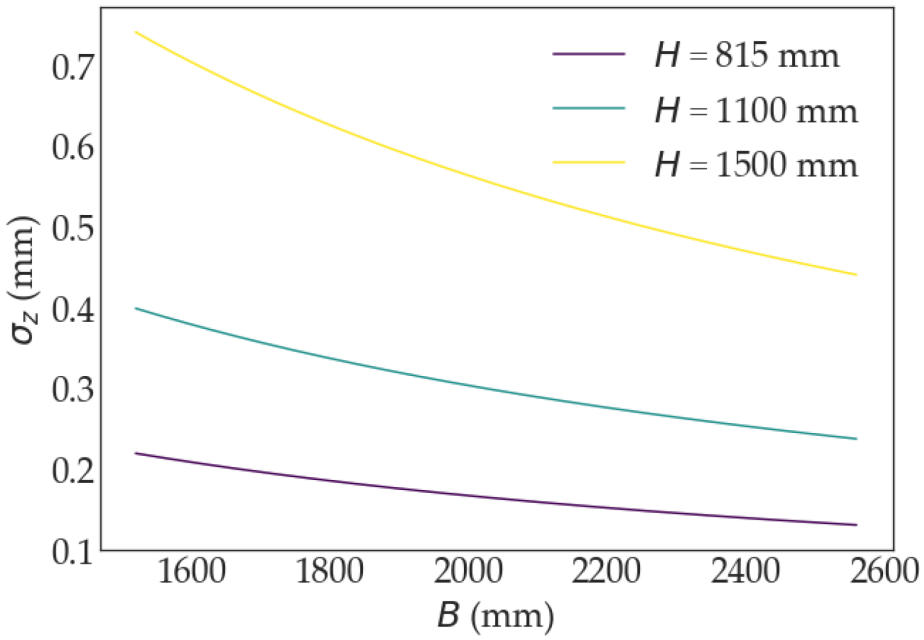

Figure 4: σ_z estimates computed by eq. (2). Colored curves refer to different D values.

eq. (2) was used to estimate the actual Z precision from the bundle adjustment solution and compare it to the expected (theoretical) precision (σ_z). To enhance the robustness of validation and test for geometrical errors, four metered tapes were placed over the bench (fig. 5), and at least nine GCPs were manually positioned throughout the scene for each acquisition date. The GCPs were positioned in a pattern to ensure a uniform distribution across the image block in both longitude and latitude. The GCPs were at three different heights: 0 m, 0.35 m, and 0.7 m.

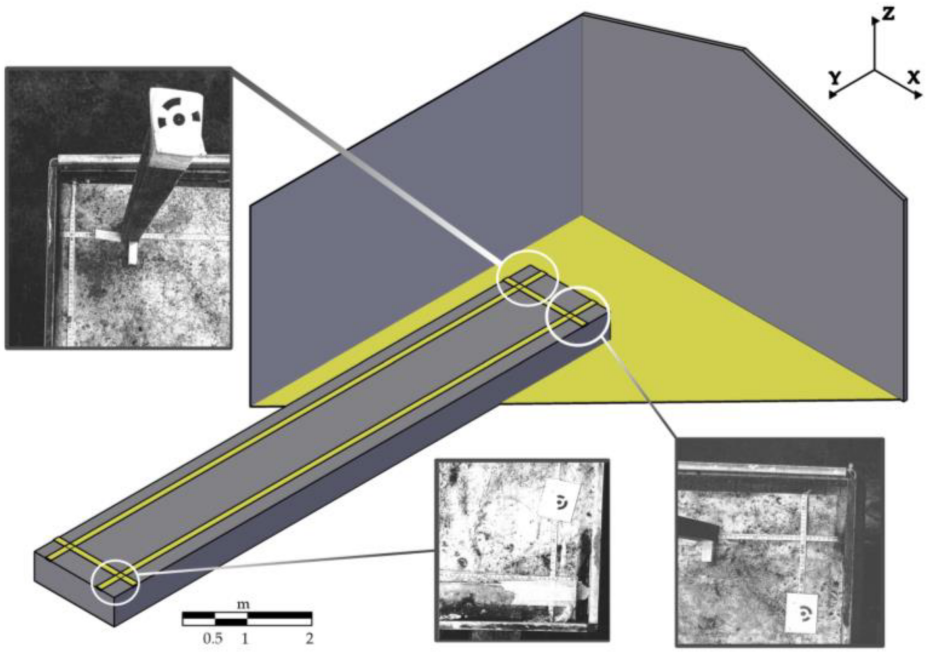

Figure 5: Metered tapes and GCPs on the bench.

In summary, the only adjustable parameters for planning the acquisition were (i) the Z position of the camera and (ii) the distance between strips (side overlap). The Z position was set at 1.5 m from the bench, and the distance between strips (on the X axis) was 20 cm in all three assessments.

2.2.2.5 Bundle Adjustment and Point Cloud Generation

Digital photogrammetry software utilize computer vision algorithms, such as the Scale-Invariant Feature Transform (SIFT), to automatically identify potential tie points in images [42, 36, 37]. Photogrammetric software may use various algorithms to match these points

across images, including Random Sample Consensus (RANSAC) or other methods, depending on computational efficiency and accuracy requirements [3]. After matching the points, software uses bundle adjustment to estimate the spatial locations of the points and the camera positions. This process takes into account the matched points and the camera's Exterior and Interior Orientation (EO/IO) parameters [34, 30, 27]. This study employed tie point identification, matching, and bundle adjustment using Agisoft Metashape version 2.1.0 (Agisoft LLC, St. Petersburg, Russia). To support image bundle adjustment, a portion of the GCPs and initial camera EO/IO parameters were provided [11, 51, 40]. As far as IO parameters are concerned, the initial values used to bootstrap adjustment were the following: (i) focal length as supplied by S3 and (ii) lens distortion parameters = 0, coordinates of the Principal Point of Autocollimation (PPA) equal to the physical center of the image (fiducial point). Sensor array and physical pixel size were set to their nominal values. The solution was spatially referenced using GCP coordinates, which are referred to as a local reference system (CRS). The resulting point cloud associates spectral values from bands to each point. These values were obtained as the mean value of the image pixels corresponding to the target points. Bundle adjustment provides estimated camera EO and IO parameters and their uncertainties, as well as all GCP coordinates estimated by the model and their corresponding errors. The GCPs involved in the bundle adjustment allow for the detection of outliers and refinement of the solution by running the bundle adjustment again after removing the outliers. To ensure accuracy, the solution was checked by three GCPs, which were not

involved in bundle adjustment. The adjustment solution was considered satisfactory if the difference between these three GCP values from the model and the reference values was less than or equal to the expected error.

2.2.2.6 Products

A digital surface model (DSM) with a GSD inherited from the previous steps was generated from the point cloud data. The DSM was then utilized to create the final multi-spectral orthomosaic (MSO) [27]. Both the DSM and MSO are projected in the CRS.

2.2.2.7 Radiometric Calibration of the Multi-Spectral Orthomosaic

MSO radiometric calibration was performed using an empirical line approach with reference reflectance values obtained from the S3 calibrated panel provided by the MAPIR company [1]. The average pixel value from each squared area of the panel having the same grey level was computed for all the bands of the non-calibrated orthomosaic. Reference reflectance values from the MAPIR calibration panel were compared with the averaged ones from the orthomosaic by scatterplot. An Ordinary Least Squares approach was used to calibrate a linear function modeling the relationship between MSO Digital Numbers and the correspondent “expected” reflectance values [15]. Calibration function definition was carried out separately for each band. The resulting functions were then applied to all the

pixels of MSO bands, resulting in a calibrated (reflectance) version of MSO fig. 6.

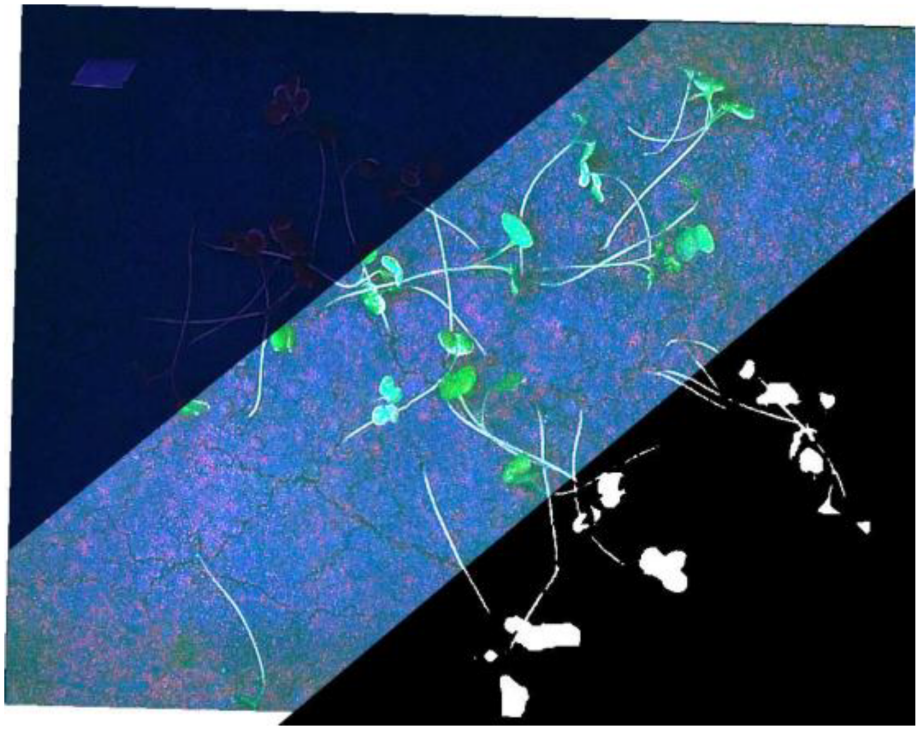

Figure 6: Non-calibrated (dark) and calibrated MSO are shown together with the last vegetation mask (white pixels) on an oil seed rape pot.

The radiometric calibration accuracy was computed as the Mean Absolute Error (MAE) between the panel ground truth values and the forecasted values [52] according to Equation

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3}$$

where y_i is the expected reflectance value of the i-calibration panel square, x_i the estimated correspondent one, and n the number of

observations.

2.2.2.8 MSO Classification

A vector format file was generated to map the area of each potted plant. A local coordinate system (CRS) was adopted. The file contains two essential pieces of information: a unique identifier for each plant and the date of assessment. The second process involved manually isolating the plant from the soil in the pot using thresholding, focusing only on the plant pixels. The soil was identified and masked by applying a bimodal threshold [43]. to the green band. The mask was then refined using a semi-automatic technique [47]. This step produced the final vegetation mask (VM) fig. 6, effectively isolating the plants for analysis.

2.2.2.9 Predictors

It is important to note that a PHYGEN estimate, in terms of a continuous variable, is the main expected outcome of this work. To achieve this task, the VM-derived area was assumed as a proxy for the Leaf Area Index (LAI). Differently, the mean (μ) and standard deviation (σ) of the following bands/indices from the calibrated S3 orthomosaic were computed: (i) Red, Green, and NIR bands, (ii) Normalized Difference Vegetation Index (NDVI), and (iii) Soil Adjusted Vegetation Index (SAVI). Additionally, the mean (μ) and standard deviation (σ) of heights of pixels belonging to VM were obtained by differencing DSM values of pixels within VM and the average of DSM values

of soil pixels. Finally, the date of acquisition (defined as DAA) was also considered to calibrate the prediction model.

Table 4: Predictors and their meanings.

Predictors	**Variables meaning**
(4) $NDVI = \frac{\rho_{NIR}-\rho_{RED}}{\rho_{NIR}+\rho_{RED}}$	where ρ_{NIR} and ρ_{RED} are the calibrated reflectance values from MSO
(5) $SAVI = \frac{1.5\cdot(\rho_{NIR}-\rho_{RED})}{(\rho_{NIR}+\rho_{RED}+0.5)}$	where ρ_{NIR} and ρ_{RED} are the calibrated reflectance values from MSO
(6) $H_P = H_V - \overline{H_S}$	where H_P is the computed pixel relative average height of the vegetation contained in a pot, H_V is the absolute height of vegetation pixel in a pot, and $\overline{H_S}$ is the average absolute height of soil level in a pot.

2.2.2.10 ML Model

The available dataset is made of 132 multivariate observations (n), each providing 14 different predictors (p). To simplify the model and reduce parameters, the least absolute shrinkage and selection operator (LASSO) model (7) was used [31]. The PHYGEN variable (y) originally expressed as a percentage, was transformed into a probability by dividing it by one hundred. As PHYGEN values range

between 0 and 1, a linear regression model is unsuitable. A logistic function was used to adjust the linear predictions from the LASSO model to the PHYGEN scale, which is relevant to human vision. Twelve variables from MSO and DAA were normalized and used as independent variables. The dataset was split into an 80% training set and a 20% testing set. A K-fold (K = 10) strategy was applied to train and cross-validate the model [33]. To ensure a balanced splitting of observations, a stratified method was used based on PHYGEN values and acquisition dates. The human visual PHYGEN was fitted using a multivariate regression model with a L_1 regularization term [23] and a least squares adjusting method. The hyperparameter λ of L_1 was determined through a cross-validation involving 5 subsets of the training data, each representing a different part of a logarithmic range developing approximately between 0.003 and 0.67. The trained model outputs were then used as inputs for a logistic function (LF) (8), which was fitted to the PHYGEN data. The function parameters were estimated using non-linear least-squares optimizatio [24, 49], with initial values inferred from the PHYGEN distribution. The optimization aimed to minimize two error functions in the model, thereby enhancing the accuracy of the PHYGEN prediction:

Table 5: Models, their loss functions, and model outputs.

Model	Loss function	Model output
LASSO	$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + L_1 = \text{min};\ L_1=\lambda\sum_{j=1}^{p}\lvert\beta_j\rvert$	$\hat{\beta}_j$, $\hat{\beta}_0$ (7)
LogisticFunction (LF)	$\sum_{i=1}^{n}\left(y_i-\frac{L}{1+e^{-k(\hat{y}_i-y_0)}}\right)^2=\text{min}$	$\hat{L}$, $\hat{k}$, $\hat{y}_0$ (8)

where y_i is the i-PHYGEN observed rate, x_ij (7) is the observed value of the j-th explaining variable, β_0 (7) is the intercept of the function, β_j (7) is the weight corresponding to j-th variable, and $\hat{y}_i$ (8) is the predicted value of the PHYGEN rate computed using weights estimated values from LASSO ($\hat{\beta}_j$,$\hat{\beta}_0$). The logistic function (8) has three parameters: L, y_0, and k. These correspond to the higher limit of the function, the inflection point of the sigmoid, and the rate of growth, respectively. The estimated values for L, k, and y_0 are, respectively, $\hat{L}$, $\hat{k}$, and $\hat{y}_0$, the correspondent estimated values. Initial values of $\hat{L}$, $\hat{k}$, and $\hat{y}_0$, needed to run the not-linear least squares were set to 100, 50, and a random value extracted in the range [0, 1], respectively.

2.2.3 Results and Discussion

2.2.3.1 Measurement Errors

The surveyed 3D coordinates of GCPs were compared to those obtained from the photogrammetric restitution of the adjusted image block to assess errors associated with geometric features. To ensure a reasonable level of robustness for the accuracy assessment despite the low number of surveyed points, a Leave One Out method was used. MAE was used as an error measure. Similarly, the accuracy of radiometric calibration was assessed using a Leave One Out (LOO) approach. An assessment was performed separately for the different dates, and the corresponding Mean Absolute Percentage Error (MAPE) values were computed. Finally, MAPE values from the different dates were averaged to define the final reference value for radiometric calibration accuracy.

2.2.3.2 Geometric Assessment Errors

Accuracy assessment concerning image block bundle adjustment was achieved at a single date level. MAE values (for each coordinate) are reported in table 6.

Table 6: XYZ errors from photogrammetric restitution in mm.

DAA	MAE$_x$ (mm)	MAE$_y$ (mm)	MAE$_z$ (mm)
3	0.57	0.61	0.62
7	0.65	0.70	0.91
14	0.67	0.68	0.89

The retained solution was deemed suitable, assuming that the differences between the main geometric features of diseased and healthy plants are greater than the reported errors. A comparison between MAE_z with the theoretical accuracy expected for the Z coordinate measure through photogrammetry eq. (2) showed that they were consistent.

2.2.3.3 Radiometric Validation

The Mean Absolute Percentage Error (MAE) of the calibration function training sample table 7 was used to estimate the goodness of function fitting.

Table 7: Radiometric Mean Absolute Percentage Error (Rad-MAPE) and ratio with the expected values obtained for the different bands and grey levels averaged along the three dates.

Band	Black (%)	Dark Gray (%)	Light Gray (%)	White (%)
Red	76.7	14.3	19.2	4.1
Green	82.8	47.5	53.2	18.1
NIR	119.6	29.2	20.7	6.2

The higher Rad-MAPE value was found for the green band, which is expected given the white balancing strategy adopted during image pre-processing. MAPE for red and NIR bands was found to be high as well, suggesting further refinements in the future to improve radiometric calibration.

2.2.3.4 Stability

The stability of the LASSO and Logistic model coefficients was analyzed. A 10-fold strategy was performed to generate an estimate for the mean and the standard deviation of coefficient estimates. fig. 7 and table 8 show related statistics.

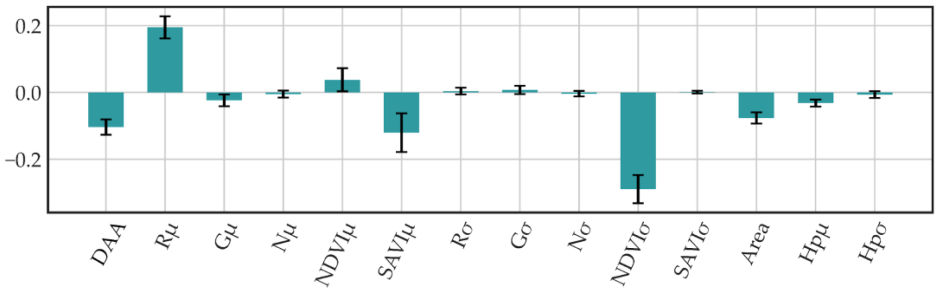

Figure 7: Mean values of LASSO β coefficients from the 10-fold approach, given for all the predictors. Whisker bars show 1-sigma LASSO β estimates.

Table 8: Mean, standard deviation, and coefficient of variation[1] values for the coefficients of the LASSO and logistic functions estimated using the 10-fold strategy.

Model	Parameter	Mean	Std. dev.	Coef.Var.[1]
LASSO	β DAA	-0.099	0.017	0.17
	β Rμ	0.2	0.05	0.25
	β SAVIμ	-0.14	0.7	0.5
	β NDVIσ	-0.28	0.04	0.14
	β Area	-0.08	0.01	0.13
	λ	0.0028	0.0013	0.46
LF	L	94.53	1.22	<0.1
	k	0.06	0.001	<0.1
	y_0	47.15	0.69	<0.1

[1] Coef.Var. is calculated with the absolute value of the mean.

nsights into the stability of the model can be gained by observing the coefficient of variation (Coef.Var.) of the most influencing parameters as estimated through the 10-fold strategy. Low values of Coef.Var. across all parameters proved that model stability is ensured. Bands and spectral indices showed the highest values of Coef.Var. This can be related to the significant uncertainty of calibrated reflectance, thus confirming the strict correspondence between measurement errors and the stability of the model (Barbedo et al. [13]).

2.2.3.5 Model Performances

Descriptive statistics of accuracy metrics were calculated with respect to the K-adjusted models used for predicting PHYGEN. MAE and the adjusted coefficient of determination $AdjR^2$ were calculated for the LASSO model, whereas the coefficient of determination R^2 was calculated for the Logistic function trained on LASSO predictions. The adjusted R^2 residual degrees of freedom were maintained equal to the number of the LASSO nonzero coefficients [54]. table 9 shows the results.

Table 9: Fit evaluation metric statistics.

Model		**MAE (PHYGEN %)**	R^2	**Adj** R^2
LASSO	Mean	11.77%	-	0.89
	Std	0.67%	-	0.03
LASSO + LF	Mean	10.66%	0.9	-
	Std	0.83%	0.03	-

The stacked model predictions ensure a mean absolute error, slightly overcoming the 11% and having a minimum coefficient of determination R_2 of about 0.9. Regarding the main goal of this work, it is worth noting that whatever the approach used to obtain an estimate of PHYGEN, its accuracy should be consistent with the one of human evaluation. According to the values reported above, the proposed method is able to provide PHYGEN scores similar to the one from experts. Our estimated accuracy (about 11%) is close to the reference threshold ordinarily accepted for PPP tests, which is 10%.

Moreover, it presents an R^2 alue similar to the SOTA model that is trained with a huge amount of data from already tested PPPs due to its CNN architecture [28]. In contrast, MAE values for PHYGEN from our model were about double that obtained from SOTA, which can exploit a huge training set more effectively. Despite this, we believe that our method is promising and affordable when considering the actual operational conditions for the estimate of PHYGEN for new and untested PPPs, which escapes from the field of application of SOTA, basing deductions on a small training set.

2.2.3.6 Compliance with ANOVA Assumptions

As previously stated, ANOVA, t-tests, and Z-tests cannot be used with ordinal discrete scale dependent variables [48]. fig. 8 shows both the ordinal discrete data used to test the model and the continuous ones from the model. This is a great improvement in the ordinary screening procedures since it enables the possibility of testing group differences through an ANOVA-based approach that a discrete variable excludes.

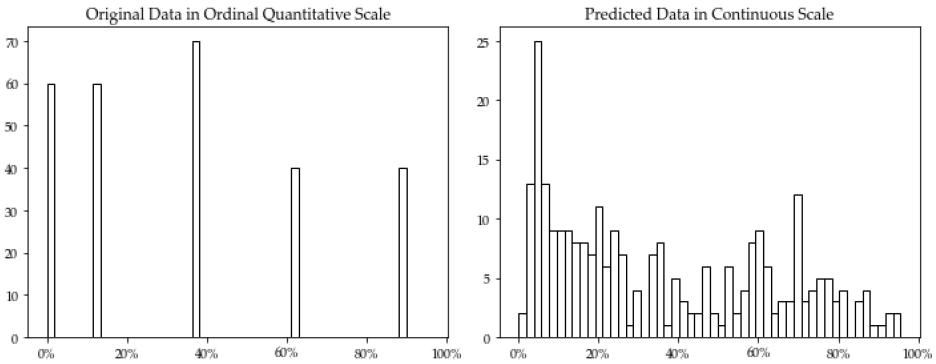

Figure 8: Discrete PHYGEN scores from the ordinary human vision-based approach (left). Continuous PHYGEN scores from the model proposed in this work (right).

2.2.4 Conclusions

The goal of this study was to test the operability and effectiveness of a controllable simple system based on multispectral digital photogrammetry and AI to support (and improve) current procedures for new PPP screening. This means that the system must be able to generate estimates of ordinarily recognized standard parameters (i.e., PHYGEN) and define the level of phytotoxicity of new PPPs before they enter the market. Basic requirements concern both compliance with accuracy standards and the robustness of the model output. The proposed method can be made operational if proper Geomatics and AI skills are properly integrated. Geomatic skills are related to proper management of the acquisition system that involves both geometric (image block bundle adjustment) and radiometric-related operations needed to prepare the data that the predictors of the PHYGEN have to be extracted from. Hardware solutions pro-

posed for the system exploit the abovementioned skills with the aim of reducing environmental and sensor-related issues. This makes acquired images more similar, partially overcoming one of the biggest problems recognized for the proper adoption of ML in phytopathometry: image features variability. A strong constraint introduced by this specific field of study is the lack of a huge training dataset that cannot be reasonably supplied for new PPPs to be screened. In such situations, this type of screening is required. The system operates in an effectively prepared greenhouse and requires significant infrastructure for the proper movement of the camera and lighting platform. In this work, we present a simple solution to these requirements. In particular, after suggesting how to pre-process the data from a photogrammetric and radiometric point of view, we found some predictors for the model to be trained that are able to exploit both the geometric and spectral content of acquired data. The predictors were analyzed and selected. They were used to train an ML algorithm integrating a LASSO and a logistic function to generate continuous estimates of PHYGEN. The robustness of the model was tested by conducting the training with a k-fold strategy and the correspondent statistics analyzed. The proposed method/system showed stability (robustness), proving to be independent of the training sample. The accuracy of PHYGEN prediction from our model is consistent with the ones from traditional methods. Compared to other AI-based approaches (i.e., SOTA), it showed slightly higher performances in terms of correlation with expert scores applied for new PPPs (our model: R^2 = 0.9, SOTA: R^2 = 0.89). In contrast, our model was not able to reach SOTA accuracy in PHYGEN scores

prediction (our model: MAE = 10.66%, SOTA: MAE = 6.74%). However, it must be noted that SOTA is not intended for predictions concerning new PPPs, and the reference values we reported refer to previously tested PPPs (i.e., providing a huge amount of training data). A surprising capability of the model was to overcome the discrete nature of expert-based scores for PHYGEN. In fact, it is able to generate continuous scores of PHYGEN, even if trained on discrete ones. Their continuous nature provides a high added value since it makes it possible to test differences among groups using ordinary ANOVA-based methods. However, some improvements are desirable, mostly in relation to a refinement of the hardware of the acquisition platform. A better-performing multispectral camera showing a higher spectral resolution and more rigorous calibration metadata is certainly a first step for future work. The active system providing controlled lighting can also be improved by using light sources that are able to generate a wider spectrum. Camera motion can be improved by using a stepper motor, allowing the possibility to stop the camera during image acquisition, thus avoiding blurring and reducing geometric deformations. Image processing could be also enhanced by strengthening automation in vegetation mask calculation from orthomosaic. The most significant improvement of the model would be to train a CNN with such a small amount of data. The final activation layer of this CNN should be set to the logistic function proposed in this work. Further studies must test data augmentation techniques and such activation layers with MAE loss to predict PHYGEN in similar setups. Regardless of the solution, we maintain that the explicability of the model, where the physical meaning of pre-

dictors and their relationships can be somehow recognized, is an added value for those applications where precise decision making is involved.

Bibliography

[1] MAPIR_Survey3_Camera_Datasheet_English.pdf. www.mapir.camera.

[2] Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. page 50.

[3] FAST APPROXIMATE NEAREST NEIGHBORS WITH AUTOMATIC ALGORITHM CONFIGURATION:. In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 331–340, Lisboa, Portugal, 2009. SciTePress - Science and and Technology Publications.

[4] Design and analysis of efficacy evaluation trials. *EPPO Bulletin*, 42(3):367–381, December 2012.

[5] PP 1/135 (4) Phytotoxicity assessment. *EPPO Bulletin*, 44(3):265–273, December 2014.

[6] PP 1/319 (1) General principles for efficacy evaluation of plant protection products with a mode of action as plant defence inducers. *EPPO Bulletin*, 51(1):5–9, April 2021.

[7] PP 1/181 (5) Conduct and reporting of efficacy evaluation trials, including good experimental practice. *EPPO Bulletin*, 52(1):4–16, 2022.

[8] Alan Agresti. Analysis of Ordinal Categorical Data.

[9] R. Alcala, H. Vitikkala, and G. Ferlet. The World Trade Organization Agreement on the Application of Sanitary and Phytosanitary Measures and veterinary control procedures. *El Acuerdo sobre la Aplicación de Medidas Sanitarias y Fitosanitarias de la Organización Mundial del Comercio y los procedimientos de control veterinario.*, 39(1):253–261, January 2020.

[10] Asif Ali, Jens C. Streibig, Joachim Duus, and Christian Andreasen. Use of Image Analysis to Assess Color Response on Plants Caused by Herbicide Application. *Weed Technology*, 27(3):604–611, 2013.

[11] Keith B. Atkinson, editor. *Close Range Photogrammetry and Machine Vision*. Whittles, Caithness, reprinted edition, 1996.

[12] Jayme G. A. Barbedo. Factors influencing the use of deep learning for plant disease recognition. *Biosystems Engineering*, 172:84–91, August 2018.

[13] Jayme G. A. Barbedo. Deep learning applied to plant pathology: The problem of data representativeness. *Tropical Plant Pathology*, 47(1):85–94, February 2022.

[14] Enrico Borgogno Mondino. Multi-temporal image co-registration improvement for a better representation and quan-

tification of risky situations: The Belvedere Glacier case study. *Geomatics, Natural Hazards and Risk*, 6(5-7):362–378, July 2015.

[15] MAPIR CAMERA. MAPIR Camera Reflectance Calibration Ground Target Package (V2). https://www.mapir.camera/products/mapir-camera-reflectance-calibration-ground-target-package-v2.

[16] Gregory A. Carter and Alan K. Knapp. Leaf optical properties in higher plants: Linking spectral characteristics to stress and chlorophyll concentration. *American Journal of Botany*, 88(4):677–684, 2001.

[17] K. S. Chiang, C. H. Bock, M. El Jarroudi, P. Delfosse, I. H. Lee, and H. I. Liu. Effects of rater bias and assessment method on disease severity estimation with regard to hypothesis testing. *Plant Pathology*, 65(4):523–535, 2016.

[18] Kuo-Szu Chiang and Clive H. Bock. Understanding the ramifications of quantitative ordinal scales on accuracy of estimates of disease severity and data analysis in plant pathology. *Tropical Plant Pathology*, 47(1):58–73, February 2022.

[19] Hangjian Chu, Chu Zhang, Mengcen Wang, Mostafa Gouda, Xinhua Wei, Yong He, and Yufei Liu. Hyperspectral imaging with shallow convolutional neural networks (SCNN) predicts the early herbicide stress in wheat cultivars. *Journal of Hazardous Materials*, 421:126706, January 2022.

[20] Samuele De Petris, Filippo Sarvia, and Enrico Borgogno-Mondino. RPAS-based photogrammetry to support tree stability assessment: Longing for precision arboriculture. *Urban Forestry & Urban Greening*, 55:126862, November 2020.

[21] EPPO. PP 1/135 (4) Phytotoxicity assessment. *EPPO Bulletin*, 44(3):265–273, 2014.

[22] 2022-06-27/29 EPPO. Digital Technology and Efficacy Evaluation of Plant Protection Products. https://www.eppo.int/MEETINGS/2022.

[23] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33:1–22, February 2010.

[24] Burton S. Garbow. MINPACK-1, Subroutine Library for Nonlinear Equation System. April 1984.

[25] David M. Gates, Harry J. Keegan, John C. Schleter, and Victor R. Weidner. Spectral Properties of Plants. *Applied Optics*, 4(1):11–20, January 1965.

[26] Sambuddha Ghosal, David Blystone, Asheesh K. Singh, Baskar Ganapathysubramanian, Arti Singh, and Soumik Sarkar. An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, 115(18):4613–4618, May 2018.

[27] Mario A. Gomarasca. Elements of Photogrammetry. In Mario A. Gomarasca, editor, *Basics of Geomatics*, pages 79–121. Springer Netherlands, Dordrecht, 2009.

[28] Laura Gómez-Zamanillo, Arantza Bereciartua-Pérez, Artzai Picón, Liliana Parra, Marian Oldenbuerger, Ramón Navarra-Mestre, Christian Klukas, Till Eggers, and Jone Echazarra. Damage assessment of soybean and redroot amaranth plants in greenhouse through biomass estimation and deep learning-based symptom classification. *Smart Agricultural Technology*, 5:100243, October 2023.

[29] Mohd Asif Hajam, Tasleem Arif, Akib Mohi Ud Din Khanday, and Mehdi Neshat. An Effective Ensemble Convolutional Learning Model with Fine-Tuning for Medicinal Plant Leaf Identification. *Information*, 14(11):618, November 2023.

[30] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, 2 edition, 2004.

[31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009.

[32] David P Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics.

[33] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learn-*

ing: With Applications in Python. Springer Texts in Statistics. Springer International Publishing, Cham, 2023.

[34] Karl Kraus. *Photogrammetry: Geometry from Images and Laser Scans*. De Gruyter, October 2011.

[35] Dawei Li, Lihong Xu, Xue-song Tang, Shaoyuan Sun, Xin Cai, and Peng Zhang. 3D Imaging of Greenhouse Plants with an Inexpensive Binocular Stereo Vision System. *Remote Sensing*, 9(5):508, May 2017.

[36] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[37] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157 vol.2, Kerkyra, Greece, 1999. IEEE.

[38] A.-K. Mahlein, M.T. Kuska, J. Behmann, G. Polder, and A. Walter. Hyperspectral Sensors and Imaging Technologies in Phytopathology: State of the Art. *Annual Review of Phytopathology*, 56(1):535–558, 2018.

[39] Anne-Katrin Mahlein. Plant Disease Detection by Imaging Sensors – Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. *Plant Disease*, 100(2):241–251, February 2016.

[40] Pierre Moulon. Positionnement robuste et précis de réseaux d□'images. page 193.

[41] B. V. Nikith, N. K. S. Keerthan, M. S. Praneeth, and Dr. T Amrita. Leaf Disease Detection and Classification. *Procedia Computer Science*, 218:291–300, January 2023.

[42] Ives Rey Otero and Mauricio Delbracio. Anatomy of the SIFT Method. *Image Processing On Line*, 4:370–396, December 2014.

[43] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979.

[44] Michael D. Owen, Damian D. Franzenburg, Dean M. Grossnickle, and James F. Lux. Evaluation of Application Timings of Warrant Herbicide for Soybean Phytotoxicity. *Iowa State University Research and Demonstration Farms Progress Reports*, 2012(1), January 2013.

[45] Françoise Petter, Anne Sophie Roy, and Ian Smith. International standards for the diagnosis of regulated pests. *European Journal of Plant Pathology*, 121(3):331–337, July 2008.

[46] Riccardo Rossi, Claudio Leolini, Sergi Costafreda-Aumedes, Luisa Leolini, Marco Bindi, Alessandro Zaldei, and Marco Moriondo. Performances Evaluation of a Low-Cost Platform for High-Resolution Plant Phenotyping. *Sensors*, 20(11):3150, January 2020.

[47] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ”GrabCut”: Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH ’04, pages 309–314, New York, NY, USA, August 2004. Association for Computing Machinery.

[48] S. S. Stevens. On the Theory of Scales of Measurement. *Science*, 103(2684):677–680, June 1946.

[49] Rainer Storn and Kenneth Price. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4):341–359, December 1997.

[50] Lijuan Tan, Jinzhu Lu, and Huanyu Jiang. Tomato Leaf Diseases Classification Based on Leaf Images: A Comparison between Classical Machine Learning and Deep Learning Methods. *AgriEngineering*, 3(3):542–558, September 2021.

[51] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883, pages 298–372. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[52] Clair Wyatt. *Radiometric Calibration: Theory and Methods*. Elsevier, December 2012.

[53] Jing Zhou, Xiuqing Fu, Leon Schumacher, and Jianfeng Zhou. Evaluating Geometric Measurement Accuracy Based on 3D

Reconstruction of Automated Imagery in a Greenhouse. *Sensors*, 18(7):2270, July 2018.

[54] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192, October 2007.

3.3 Binary and Nominal Variables: Anomaly Detection for Plant Deseases Classification

This study investigates the application of unsupervised learning approaches for binary and nominal variable classification in disease detection. The research demonstrates how pre-trained models can be leveraged for plant health assessment without task-specific training, using anomaly detection techniques to classify healthy versus diseased plants within the EPPO framework.

Abstract

This study systematically evaluates the efficacy of 56 pretrained neural network architectures, used without fine-tuning, as feature extractors for plant disease anomaly detection across laboratory and field environments. We compare convolutional and transformer-based networks in conjunction with various dimensionality reduction techniques and anomaly detection algorithms to address the performance gap between controlled and real-world imaging conditions. Using apple leaf disease datasets (Plant Village and Plant Pathology) containing identical disease classes, we implement two complementary evaluation strategies: anomaly detection trained solely on healthy samples and clustering-based classification to distinguish between specific disease types. Results reveal a consistent 5-10% performance reduction when transitioning from laboratory to field images, highlighting the challenge of developing robust field-deployable systems. The lightweight ShuffleNet_v2_x1_0 architecture (2.3M parameters) outperformed substantially larger models like DINOv2 (300M) and ViT (86M) in field conditions, challenging the assumption that larger models necessarily yield better performance for specialized tasks. Among dimensionality reduction techniques, t-SNE consistently outperformed others, while Local Outlier Factor demonstrated the most stable anomaly detection performance across datasets. For clustering, density-based DBSCAN with ShuffleNet_v2_x1_0 achieved

superior performance on field images. These findings provide practical insights for developing computationally efficient plant disease detection systems for resource-constrained environments, demonstrating that anomaly detection approaches with off-the-shelf pretrained models offer viable alternatives to supervised classification, especially when comprehensive labeled datasets are impractical.

2.3.1 Introduction

Plant diseases pose a significant threat to agricultural productivity, food security, and economic stability worldwide, with estimated global crop losses exceeding 20-40% annually due to pathogens [19]. Early and accurate detection of plant diseases is crucial for implementing timely interventions, reducing pesticide use, and preventing disease spread across agricultural landscapes [15]. Traditional disease detection methods rely heavily on visual inspection by trained experts, which is time-consuming, labor-intensive, and subject to human error [3].

In recent years, advances in computer vision and machine learning have enabled automated approaches to plant disease detection, offering the potential for more scalable, consistent, and objective diagnostic capabilities [8, 16]. Deep learning approaches, particularly convolutional neural networks (CNNs) and vision transformers, have demonstrated remarkable success in classifying plant diseases from leaf images [22]. However, these supervised approaches require large amounts of labeled training data for each disease class, which is often impractical to obtain for the diverse range of plant pathogens and their varying manifestations [26].

Anomaly detection presents a promising alternative paradigm that requires training only on healthy samples, identifying diseased specimens as deviations from the normal state [18, 6]. This approach aligns well with agricultural monitoring scenarios where healthy plants constitute the majority class, and various diseases represent anoma-

lous conditions [12]. Additionally, anomaly detection frameworks can potentially identify novel or previously unseen disease manifestations that supervised classifiers would struggle to recognize [5].

A critical challenge in developing robust plant disease detection systems is the significant performance gap between controlled laboratory environments and real-world field conditions [25]. Laboratory-acquired images typically feature isolated leaves against uniform backgrounds with consistent lighting, while field-acquired images contain variable illumination, complex backgrounds, and diverse perspectives that can dramatically affect feature extraction and classification performance [3].

This study addresses these challenges by comprehensively evaluating the efficacy of various neural network architectures as feature extractors for anomaly detection across both laboratory and field-acquired apple leaf disease datasets. By systematically comparing convolutional and transformer-based networks in conjunction with different dimensionality reduction techniques and anomaly detection algorithms, we aim to identify robust methodologies that can translate from controlled environments to practical field applications.

Our work makes several key contributions: (1) a systematic evaluation of 56 neural network architectures as feature extractors for plant disease anomaly detection; (2) comparative analysis of performance across laboratory and field imaging conditions using parallel datasets with matching disease classes; (3) identification of lightweight models that achieve benchmark accuracy while minimizing computational requirements; and (4) practical insights into the

most effective combinations of feature extraction, dimensionality reduction, and anomaly detection approaches for agricultural disease monitoring applications.

2.3.2 Materials and Method

2.3.2.1 Dataset

This study utilized two complementary apple leaf disease datasets to evaluate the robustness of feature extraction methods across different imaging conditions. The datasets represent controlled laboratory conditions and natural field environments, respectively, while containing the same disease classes.

The Plant Village dataset [11] consists of laboratory-acquired images of individual plant leaves photographed against controlled backgrounds. For this study, the apple leaf subset was utilized, which includes segmented images of single leaves. These images were captured under consistent lighting conditions with uniform backgrounds, resulting in standardized image dimensions of 256×256 pixels.

The Plant Pathology dataset [24], collected as part of a Kaggle competition, contains field-acquired images of apple leaves in their natural environment. Unlike the controlled Plant Village images, these photographs exhibit varying lighting conditions, backgrounds, perspectives, and image dimensions. The images capture leaves still attached to the tree or branch, providing a more challenging and realistic scenario for disease detection.

To enable fair comparison between the datasets, several preprocessing steps were implemented. First, the datasets were balanced to contain identical disease classes (healthy, apple scab, and cedar apple rust). Second, the number of samples per class was standardized by removing excess observations from either dataset where necessary. Both datasets were then processed to ensure consistent sample counts while maintaining their inherent characteristics regarding acquisition conditions.

Table 1: Summary of the standardized apple leaf disease datasets

Dataset	Class	Samples	Image Size	Acquisition
Plant Village	Healthy	516	256×256	Laboratory
	Cedar apple rust	275		
	Apple scab	583		
Plant Pathology	Healthy	516	Variable	Field
	Cedar apple rust	275		
	Apple scab	583		

The dataset combination provides an opportunity to assess how feature extraction methods perform across different imaging conditions while maintaining consistent disease classes. The laboratory-acquired Plant Village images offer an idealized, controlled scenario, while the field-acquired Plant Pathology images present a more challenging real-world testing environment.

Figure 1 illustrates representative samples from both datasets, clearly showing the substantial differences in imaging conditions between laboratory and field acquisitions. The controlled Plant Village images feature isolated leaves against uniform backgrounds with consistent lighting, while the Plant Pathology images exhibit natural field

conditions with variable lighting, complex backgrounds, and diverse perspectives.

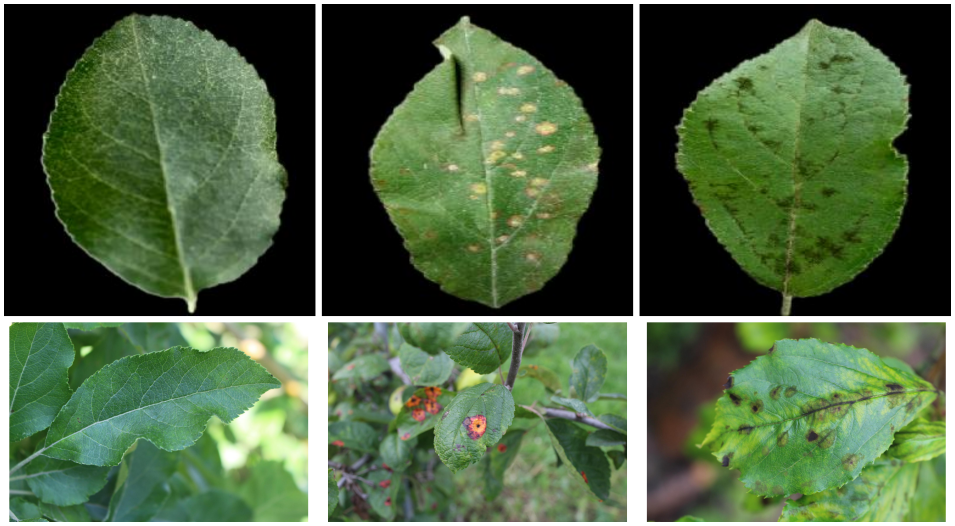

Figure 1: Representative examples from the apple leaf disease datasets used in this study. Top row shows Plant Village dataset samples: healthy leaf (left), cedar apple rust (center), and apple scab (right). Bottom row shows Plant Pathology dataset samples: healthy leaf (left), cedar apple rust (center), and apple scab (right). Note the controlled laboratory conditions with uniform backgrounds in Plant Village images versus the natural field conditions with variable lighting and complex backgrounds in Plant Pathology images.

2.3.2.2 Tested Backbones

This study evaluates a comprehensive set of neural network architectures as feature extractors for anomaly detection. We implemented both convolutional neural networks (CNNs) and transformer-based architectures pre-trained on ImageNet to extract meaningful representations from input images.

Table 2: Overview of neural network backbone architectures evaluated in this study

Backbone	Param (M)	Input Size	Backbone	Param (M)	Input Size
densenet121	8.0	224×224	regnet_y_8gf	39.4	224×224
densenet161	28.7	224×224	resnet101	44.5	224×224
densenet169	14.1	224×224	resnet152	60.2	224×224
densenet201	20.0	224×224	resnet18	11.7	224×224
dinov2_vitb14	86.0	224×224	resnet34	21.8	224×224
dinov2_vitl14	300.0	224×224	resnet50	25.6	224×224
dinov2_vits14	21.0	224×224	resnext101_32x8d	88.8	224×224
googlenet	13.0	224×224	resnext101_64x4d	83.5	224×224
inception_v3	27.2	299×299	resnext50_32x4d	25.0	224×224
mobilenet_v3_large	5.5	224×224	shufflenet_v2_x0_5	1.4	224×224
mobilenet_v3_small	2.5	224×224	shufflenet_v2_x1_0	2.3	224×224
regnet_x_16gf	54.3	224×224	shufflenet_v2_x1_5	3.5	224×224
regnet_x_1_6gf	9.2	224×224	shufflenet_v2_x2_0	7.4	224×224
regnet_x_32gf	107.8	224×224	swin_b	87.8	224×224
regnet_x_3_2gf	15.3	224×224	swin_s	49.6	224×224
regnet_x_400mf	5.5	224×224	swin_t	28.3	224×224
regnet_x_800mf	7.3	224×224	swin_v2_b	87.9	224×224
regnet_x_8gf	39.6	224×224	swin_v2_s	49.7	224×224
regnet_y_16gf	83.6	224×224	swin_v2_t	28.4	224×224
regnet_y_1_6gf	11.2	224×224	vgg11	132.9	224×224
regnet_y_32gf	145.0	224×224	vgg11_bn	132.9	224×224
regnet_y_3_2gf	19.4	224×224	vgg13	133.0	224×224
regnet_y_400mf	4.3	224×224	vgg13_bn	133.0	224×224
regnet_y_800mf	6.4	224×224	vgg16	138.4	224×224
vgg16_bn	138.4	224×224	vit_l_16	304.3	224×224
vgg19	143.7	224×224	vit_l_32	306.5	224×224
vgg19_bn	143.7	224×224	wide_resnet101_2	126.9	224×224
vit_b_16	86.6	224×224	wide_resnet50_2	68.9	224×224

Convolutional Neural Networks

We investigated several CNN architecture families:

- **ResNet family:** ResNet18, ResNet34, ResNet50, ResNet101,

ResNet152, which utilize residual connections to enable training of deeper networks [9]. Additionally, we included variants with wider channels (Wide ResNet50, Wide ResNet101) and grouped convolutions (ResNeXt50, ResNeXt101).

- **VGG family:** VGG11, VGG13, VGG16, VGG19, and their batch-normalized counterparts, representing traditional deep CNN architectures with sequential convolutional layers [21].
- **DenseNet family:** DenseNet121, DenseNet161, DenseNet169, DenseNet201, featuring dense connectivity patterns that strengthen feature propagation [10].
- **Efficient architectures:** EfficientNet (B0-B7), EfficientNetV2 (S, M, L), MobileNetV2, MobileNetV3, which are optimized for computational efficiency while maintaining high accuracy [23].
- **Other CNN architectures:** GoogleNet, Inception-v3, RegNet, ShuffleNet, and SqueezeNet variations, each with unique architectural innovations designed to improve performance or efficiency.

Transformer-based Architectures

We also examined vision transformers that have demonstrated strong performance in recent years:

- **Vision Transformer (ViT):** ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, ViT-H/14, which apply the transformer architecture directly to image patches [7].

- **Swin Transformer:** Swin-T, Swin-S, Swin-B and their V2 variants, which incorporate hierarchical feature maps and shifted windows for more efficient attention computation [14].
- **DINOv2:** DINOv2-ViT-S/14, DINOv2-ViT-B/14, DINOv2-ViT-L/14, which are self-supervised vision transformers trained using distillation with no labels [17].

Feature Extraction Methodology

For all architectures, we removed the classification heads and extracted features from the penultimate layer. For CNNs, this typically corresponds to the output after global average pooling, while for transformers, we used the [CLS] token representation. All models were pre-trained on ImageNet and used without fine-tuning to evaluate their transfer learning capabilities for anomaly detection.

For standard torchvision models, we utilized the official pre-trained weights [1]. For DINOv2 models, we loaded weights directly from the official Facebook Research repository [2]. Input images were processed using the standard preprocessing pipeline recommended for each model, including resizing, normalization, and in some cases, center cropping.

2.3.2.3 Evaluation Strategies

We implemented two complementary strategies to evaluate the efficacy of extracted features:

Anomaly Detection Approach

The extracted features were used as input to anomaly detection algorithms. For each dataset, these algorithms were trained using only the healthy samples and evaluated on the diseased samples within the same dataset. This approach allowed us to assess how well the feature extractors could separate normal from anomalous samples across different imaging conditions.

Clustering-based Classification

We also evaluated whether the dimensionality-reduced features preserved sufficient class-discriminative information for conventional clustering algorithms to recover the original disease classes. This approach differs from anomaly detection by attempting to distinguish between specific disease types rather than just identifying abnormalities.

We tested multiple clustering algorithms:

- **K-Means:** A centroid-based algorithm that partitions the data into k clusters, with each observation belonging to the cluster with the nearest mean.
- **Hierarchical Clustering:** An agglomerative approach that builds nested clusters by merging or splitting them successively.
- **Gaussian Mixture Models:** A probabilistic model that assumes data points are generated from a mixture of several Gaussian distributions.

- **DBSCAN:** A density-based clustering algorithm that groups together points that are closely packed in feature space.

To evaluate clustering performance, we mapped each cluster to its most common ground truth label and calculated Cohen's Kappa coefficient to measure the agreement between clustering assignments and original disease classifications.

2.3.2.4 Dimensionality Reduction

To visualize the extracted features and assess their separability, we applied dimensionality reduction techniques. We selected t-SNE, UMAP, and PCA for this purpose:

- **t-SNE (t-distributed Stochastic Neighbor Embedding):** A non-linear dimensionality reduction technique that is particularly effective for visualizing high-dimensional data in lower dimensions (typically 2D or 3D). It focuses on preserving local structures and is widely used for visualizing clusters in feature spaces.
- **UMAP (Uniform Manifold Approximation and Projection):** A manifold learning technique that preserves both local and global structures in the data. UMAP is often faster than t-SNE and can produce more interpretable embeddings, making it suitable for visualizing complex datasets.
- **PCA (Principal Component Analysis):** A linear dimensionality reduction method that transforms the data into a new co-

ordinate system, where the greatest variance lies on the first coordinates (principal components). PCA is computationally efficient and provides a global view of the data structure.

These techniques were applied to the extracted features from both datasets, allowing us to visualize the distribution of healthy and diseased samples in lower-dimensional spaces. The visualizations provided insights into the separability of different classes and the effectiveness of the feature extractors in capturing relevant information for anomaly detection.

2.3.2.5 Anomaly Detection Algorithms

We implemented a range of anomaly detection algorithms to evaluate the performance of the extracted features. The algorithms were selected based on their popularity and effectiveness in various domains, including:

Statistical Methods

- **IQR with Confidence Interval:** This approach combines robust statistics with probabilistic bounds. First, we use the interquartile range (IQR) of healthy samples to identify potential outliers. We then calculate a confidence interval (95%) around the mean of the remaining inliers. Any sample falling outside this interval is classified as anomalous.

Machine Learning Methods

- **Isolation Forest:** This algorithm [13] isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Anomalies require fewer partitions to be isolated, resulting in shorter average path lengths. We used a contamination parameter of 0.1 for all experiments.

- **One-Class SVM:** This method [20] learns a boundary around normal data points in feature space. Samples outside this boundary are classified as anomalies. We employed the RBF kernel with nu=0.1, training only on healthy samples.

- **Local Outlier Factor (LOF):** This density-based algorithm [4] compares the local density of a point with the densities of its neighbors. Points with substantially lower density than their neighbors are considered anomalies. We configured LOF in novelty mode with optimal neighborhood size.

- **Gaussian Mixture Model (GMM):** This probabilistic model assumes that normal data points are generated from a mixture of Gaussian distributions. We fit a single-component GMM to healthy samples and identified anomalies as points with low probability density, using the 1st percentile of healthy samples' scores as the threshold.

Experimental Setup

We implemented two parallel experimental workflows to evaluate the feature representations:

Anomaly Detection Workflow

For evaluating anomaly detection capabilities, we followed this procedure for each dataset independently:

1. Extract features from the dataset using the backbone networks.
2. Apply dimensionality reduction techniques (t-SNE, UMAP, or PCA) to the extracted features.
3. Train the anomaly detection algorithms using only the healthy samples from the dataset.
4. Evaluate performance on the diseased samples from the same dataset, where all non-healthy classes should be detected as anomalies.

We assessed each algorithm using standard binary classification metrics: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

Clustering-based Classification Workflow

For evaluating the class-discriminative information in the feature space, we implemented:

1. Extract features from the dataset using the backbone networks.
2. Apply dimensionality reduction techniques to reduce feature dimensionality.
3. Apply various clustering algorithms (K-Means, Hierarchical Clustering, GMM, DBSCAN) to the reduced features.

4. Map each resulting cluster to the most common ground truth label among its members.

5. Calculate Cohen's Kappa coefficient between the cluster assignments and the original disease classifications to measure agreement beyond chance.

The clustering approach provides insights into whether the feature extractors capture sufficient information to distinguish between specific disease classes, rather than just separating normal from abnormal samples. It also serves as a more challenging evaluation scenario that mimics unsupervised disease classification.

By applying both methodologies to the controlled laboratory images (Plant Village) and the variable field images (Plant Pathology), we could comprehensively evaluate how different feature extractors perform across varying imaging conditions, which is crucial for practical disease detection applications in agriculture.

2.3.3 Results and Discussion

Our analysis of anomaly detection and clustering performance across different backbone architectures, dimensionality reduction techniques, and anomaly detection algorithms revealed several combination achiving the accuracy benchmark. Figure 2 and Figure 3 show the distribution of anomaly detection accuracy across all tested configurations respectively for the Plant Village and the Plant Pathology datasets. In the same way, Figure 4 and Figure 5 show the dis-

tribution of clustering accuracy across all tested configurations respectively for the Plant Village and the Plant Pathology datasets.

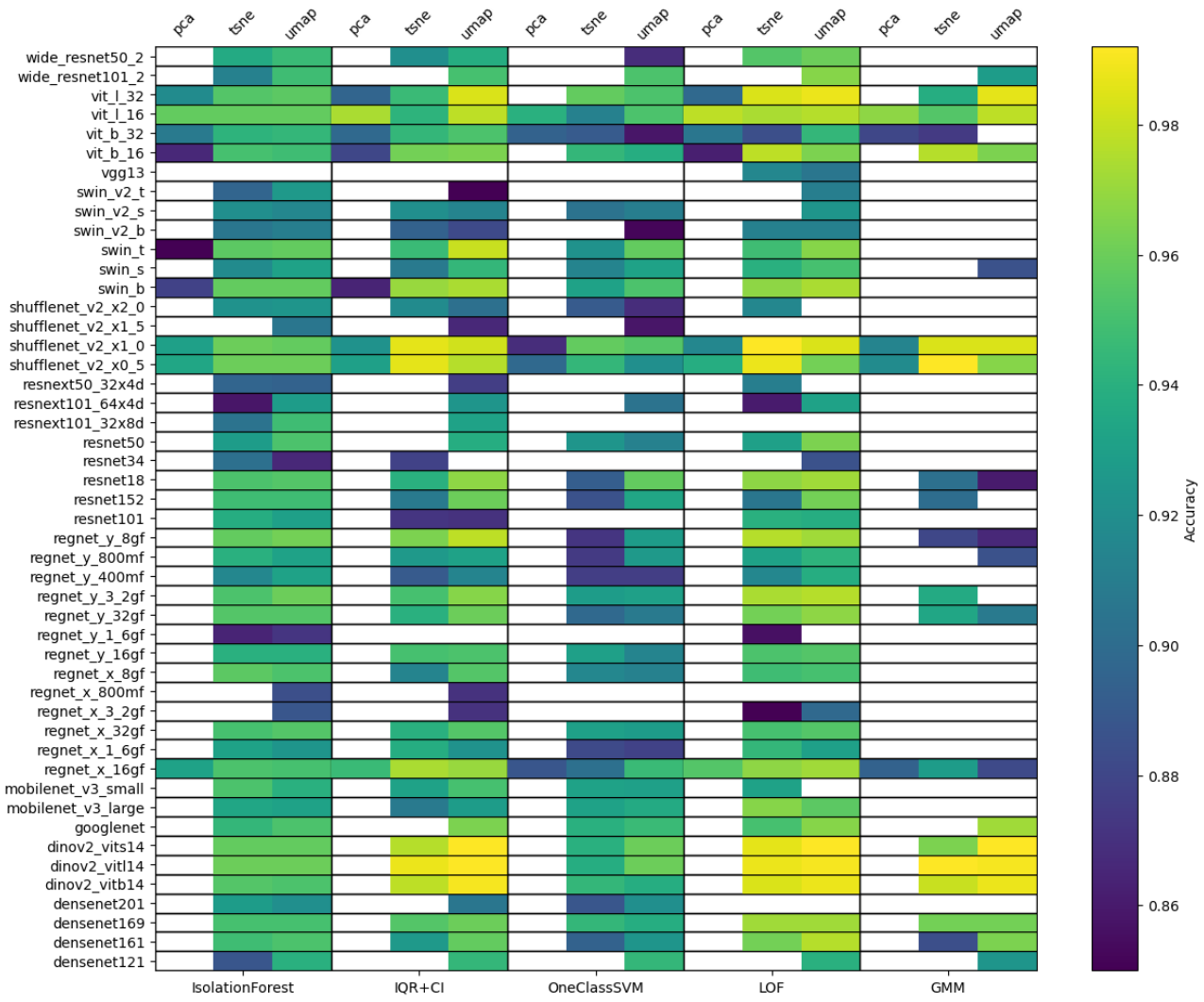

Figure 2: Anomaly detection performance across different backbone architectures and dimensionality reduction techniques on the Plant Village dataset. Backbones on y-axis, anomaly detetection algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for each backbone-detection-reduction combination.

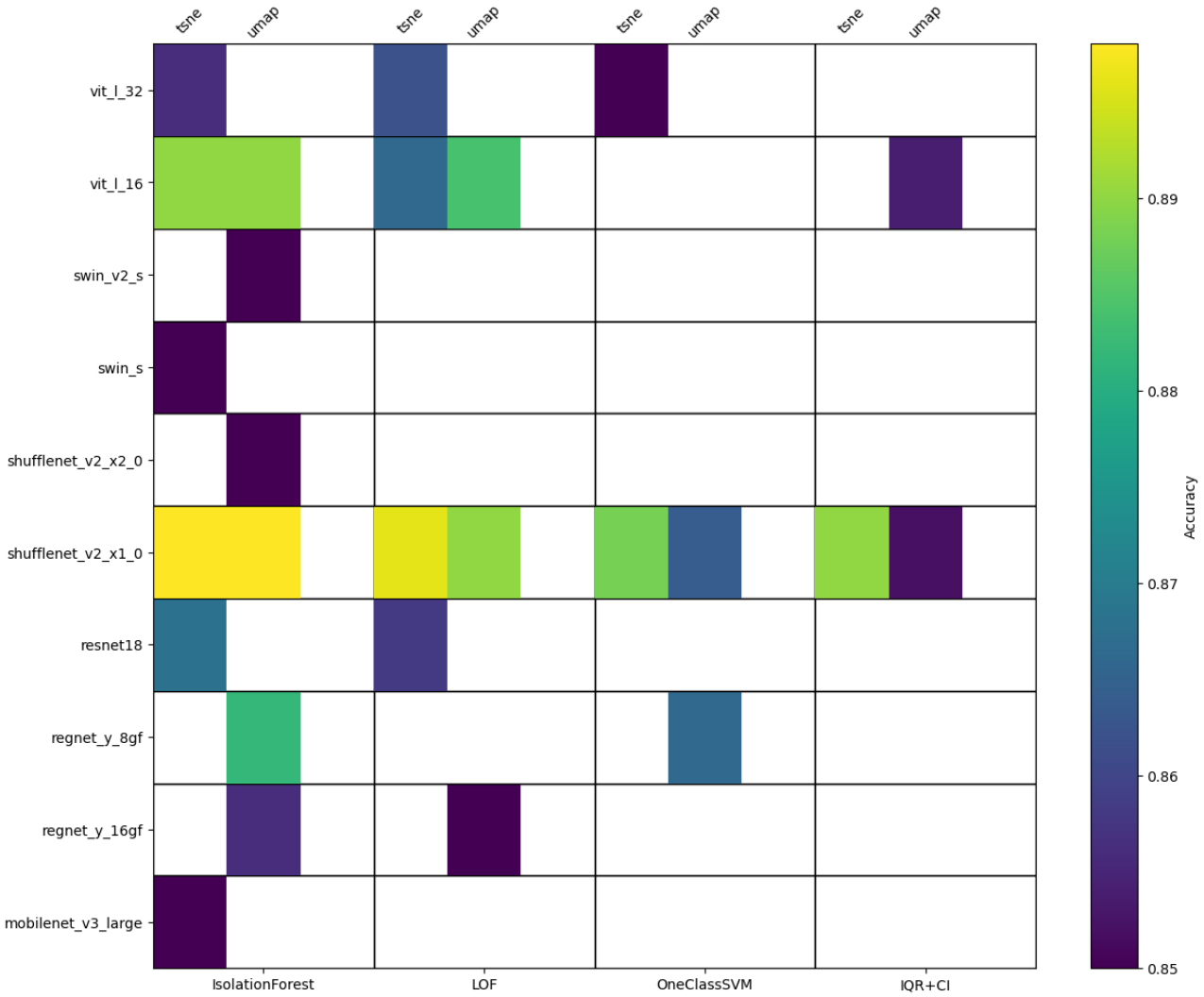

Figure 3: Anomaly detection performance across different backbone architectures and dimensionality reduction techniques on the Plant Pathology dataset. Backbones on y-axis, anomaly detetection algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for each backbone-detection-reduction combination.

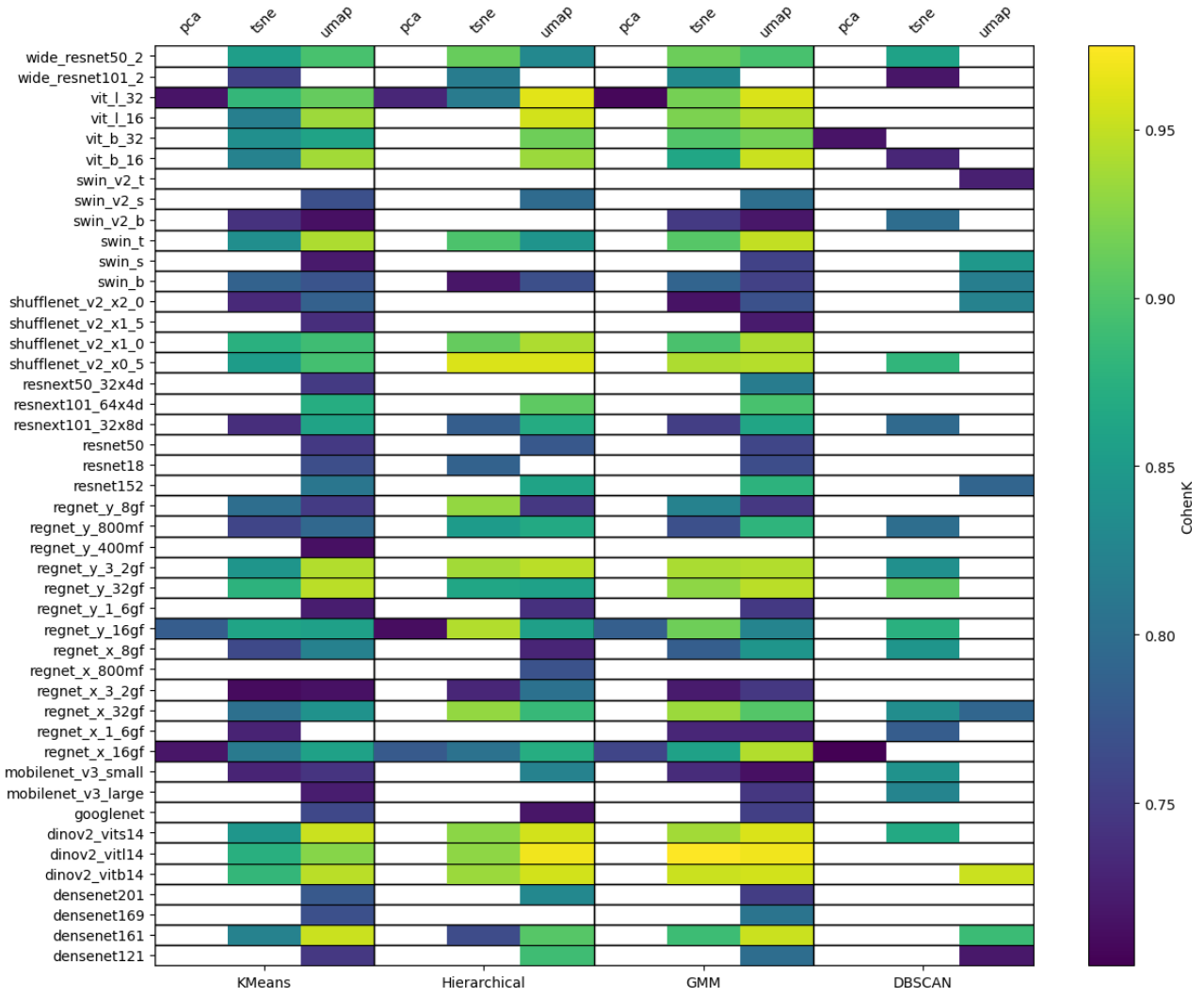

Figure 4: Clustering performance across different backbone architectures and dimensionality reduction techniques on the Plant Village dataset. Backbones on y-axis, clustering algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for each backbone-detection-reduction combination.

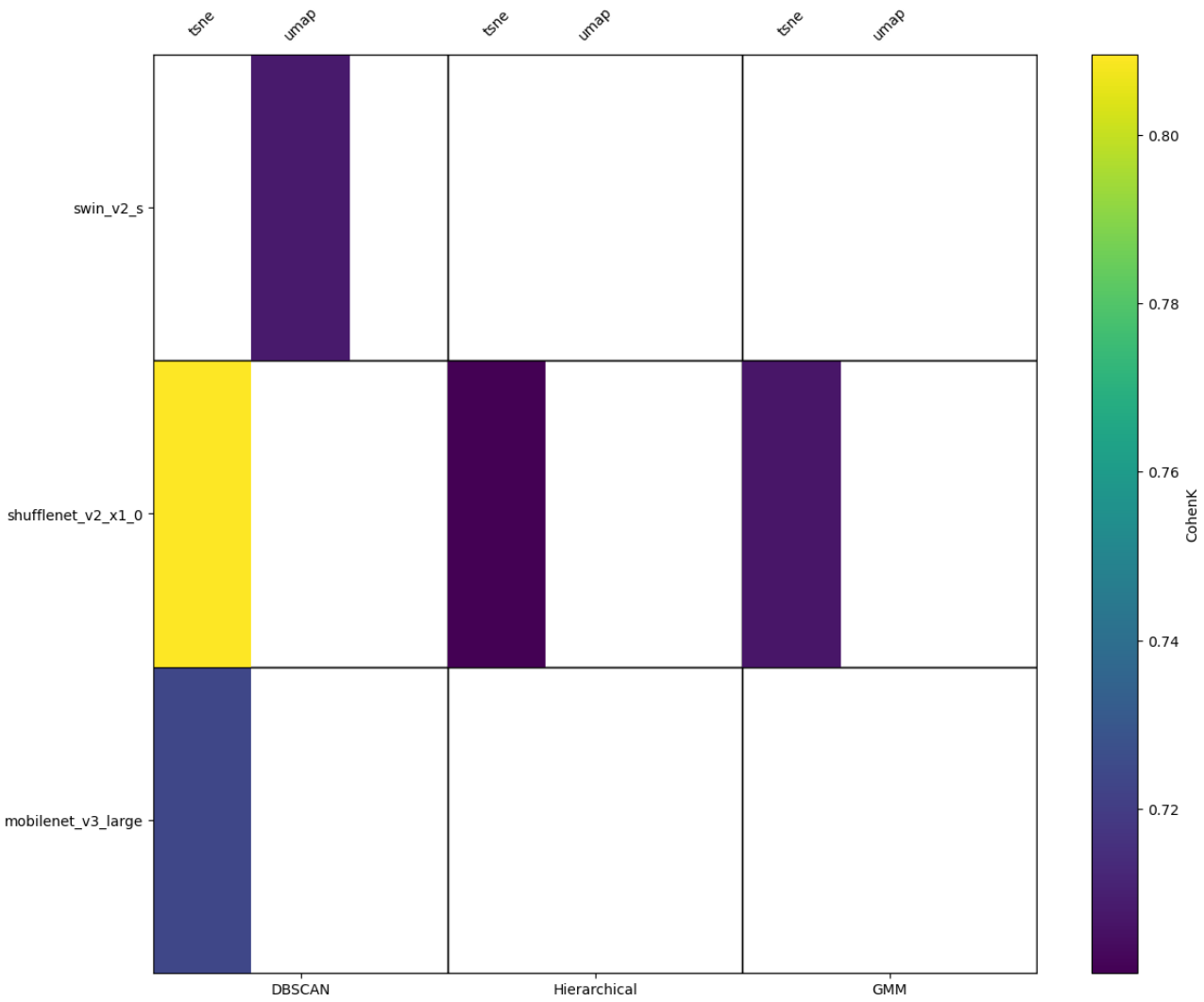

Figure 5: Clustering performance across different backbone architectures and dimensionality reduction techniques on the Plant Pathology dataset. Backbones on y-axis, clustering algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for each backbone-detection-reduction combination.

Dataset related performances

The analysis of performance metrics across datasets revealed sig-

nificant differences in model efficacy between laboratory-acquired and field-acquired images. The Plant Village dataset, consisting of controlled laboratory images with segmented leaves against uniform backgrounds, consistently enabled superior performance compared to the Plant Pathology dataset across all evaluation metrics.

For anomaly detection tasks, accuracy scores on the Plant Village dataset frequently exceeded 90%, as shown in Figure 2. In contrast, the same architectures applied to the Plant Pathology dataset did not achieved the benchmark or got a performance decrease of approximately 5-10%, as illustrated in Figure 3.

This performance gap was even more pronounced in clustering tasks, where Cohen's Kappa coefficients reached as high as 0.9 with 12 backbones on Plant Village data but peaked at 0.80 on Plant Pathology data and achived benchmark with only three backbones, as seen in Figures 4 and 5. The controlled environment and object-focused nature of the Plant Village dataset allowed models to concentrate on leaf characteristics directly relevant to disease detection without the confounding variables present in field conditions.

The Plant Pathology dataset's variable lighting, complex backgrounds, and inconsistent perspectives presented a substantially more challenging scenario for feature extraction focusing on diseseas sympthoms.

These findings highlight the significant challenge of transitioning plant disease detection systems from controlled laboratory environments to real-world field applications, where background elements and environmental variability introduce substantial complexity to the feature extraction and classification process.

Nethertheless, the benchmark was achieved also for the challenging field-acquired images, indicating that the selected architectures and methods are capable of generalizing to real-world conditions.

Effect of Feature Extraction Architecture

Among the tested architectures, the ShuffleNet_v2 feature extractor consistently demonstrated strong performance across both datasets, with the x1_0 size version achieving the benchmark for both anomaly and clusterization tasks. Other well performing architectures were DINOv2 and ViT on Plant Village dataset for both tasks, but they did not perform consistently on the Plant Pathology dataset. Remarkably ShuffleNet_v2_x1_0 is a lightweight architecture with only 2.3M parameters, in respect to the DINOv2 and ViT architectures which have 300M and 86M parameters respectively. This suggests that the selected feature extractors are capable of achieving high performance even with limited computational resources, making them suitable for deployment in resource-constrained environments.

Impact of Dimensionality Reduction

Different dimensionality reduction techniques showed varying effectiveness:

- **t-SNE** consistently yielded the highest performances on both tasks and datasets when in combination with ShuffleNet_v2_x1_0.
- **UMAP** performed competitively with t-SNE, in some cases surpassing it, but not the best option for all tasks and datasets.
- **PCA**, while computationally efficient, generally produced lower accuracy compared to t-SNE and UMAP, indicating that lin-

ear dimensionality reduction may not sufficiently preserve the complex structure necessary for plant disease detection.

The choice of dimensionality reduction technique significantly influenced the performance of both anomaly detection and clustering tasks. t-SNE and UMAP were particularly effective in preserving the local structure of the data, leading to better separability of classes in the reduced feature space.

Comparison of Anomaly Detection Algorithms

Among the tested anomaly detection algorithms:

- **Isolation Forest** exalled on Plant Pathology dataset while the performances with Plant Village were low in respect to the others methods.
- **One-Class SVM** did not exalled on any of the datasets.
- **LOF** showed the most stable performances.
- **GMM** demonstrated comparable performance with LOF on Plant Village, but it never reached the benchmark on Plant Pathology.
- **IQR with Confidence Interval**, while simpler than the machine learning approaches, still achieved respectable performance on both datasets, highlighting the effectiveness of statistical approaches for this task.

The Plant Pathology dataset generally yielded lower performance

compared to Plant Village across all algorithms, reflecting the greater difficulty of analyzing field-acquired images with variable conditions.

Clustering Performance

The clustering-based classification approach revealed complementary insights about the discriminative power of extracted features.

- **K-Means, Hierarchical, and GMM** Good results on Plant Village achieving benchmark with multiple backbones, but on Plant Pathology only K-Means and GMM reached the benchmark.
- **DBSCAN** achieved the benchmark on Plant Village with less backbones in respect the other methods, while on Plant Pathology achieved the best result with ShuffleNet_v2_x1_0.

The clustering-based classification approach revealed complementary insights about the discriminative power of extracted features, with significant differences in algorithm performance across datasets.

- **K-Means, Hierarchical, and GMM** achieved strong results on Plant Village, reaching the benchmark with multiple backbone architectures. However, on the more challenging Plant Pathology dataset, only K-Means and GMM reached the benchmark. This suggests that centroid and distribution-based approaches perform consistently when the number of clusters is explicitly defined to match the disease classes.
- **DBSCAN** exhibited a distinct behavior pattern, achieving the benchmark on Plant Village with fewer backbones compared

to other methods, while on Plant Pathology it achieved the best overall result specifically with ShuffleNet_v2_x1_0. This unique performance profile can be attributed to DBSCAN's density-based approach, which differs fundamentally from the other algorithms in several ways:

- Unlike parametric methods that assume specific cluster shapes, DBSCAN identifies arbitrarily shaped clusters based on density variations, potentially capturing the complex symptom patterns in field conditions more effectively.
- DBSCAN automatically estimates its critical epsilon parameter based on the nearest neighbor distances in the feature space, making it particularly responsive to the actual distribution characteristics rather than prior assumptions.
- Its built-in outlier detection capability, which labels points in low-density regions as noise, provides natural robustness against the variable imaging conditions present in the Plant Pathology dataset.
- The exceptional performance with ShuffleNet_v2_x1_0 suggests this lightweight architecture (2.3M parameters) produces feature distributions with clearer density gradients between disease classes, despite having significantly fewer parameters than transformer-based alternatives.

These findings indicate that while conventional clustering methods perform well in controlled environments, density-based approaches

may offer advantages for disease detection in variable field conditions, particularly when paired with efficient feature extractors that create well-separated density regions in the feature space.

2.3.4 Conclusions

This comprehensive evaluation of neural network architectures as feature extractors for plant disease anomaly detection has yielded several important findings with significant implications for agricultural monitoring applications.

Our first key finding revealed a consistent performance gap between laboratory and field-acquired images, with detection accuracy typically 5-10% lower on field images. This quantifies the substantial challenge of translating plant disease detection systems from controlled environments to practical field applications. Despite this gap, our study identified combinations of feature extractors and detection algorithms that achieved benchmark performance even in challenging field conditions, demonstrating that robust field-deployable systems are achievable.

The ShuffleNet_v2_x1_0 architecture emerged as the most consistently effective feature extractor across both datasets and evaluation methodologies. Remarkably, this lightweight network (2.3M parameters) outperformed substantially larger models like DINOv2 (300M parameters) and ViT (86M parameters) in field conditions. This finding challenges the common assumption that larger, more complex models necessarily yield better performance for specialized tasks.

Instead, it suggests that computational efficiency and targeted feature extraction may be more valuable than model capacity for plant disease detection, particularly in resource-constrained deployment scenarios.

Among dimensionality reduction techniques, t-SNE consistently yielded the highest performance across most configurations, with UMAP following closely. The substantially lower performance of PCA indicates that nonlinear dimensionality reduction techniques better preserve the complex feature relationships crucial for disease differentiation. This finding highlights the importance of maintaining local neighborhood structures in the reduced feature space for effective anomaly detection.

The comparison of anomaly detection algorithms revealed that LOF demonstrated the most stable performance across datasets, while Isolation Forest excelled specifically on field-acquired images. This suggests that different detection methodologies have complementary strengths depending on image acquisition conditions. For practical field applications, ensemble approaches combining multiple detection algorithms might prove beneficial.

For clustering-based classification, DBSCAN with ShuffleNet_v2_x1_0 achieved superior performance on field images compared to other combinations. This density-based approach appears particularly well-suited to handling the variable imaging conditions present in field settings, capturing the natural density variations between healthy and diseased samples in the feature space.

These findings have important practical implications for agricultural

disease monitoring systems. By selecting lightweight, efficient architectures like ShuffleNet_v2_x1_0, developers can create deployment-ready solutions for resource-constrained environments such as edge devices or mobile applications. The established benchmark performance on field-acquired images demonstrates that anomaly detection approaches are viable alternatives to supervised classification, especially in scenarios where obtaining comprehensive labeled datasets for every potential disease is impractical.

Future research directions should explore fine-tuning strategies specifically for agricultural domain adaptation, which may further close the performance gap between laboratory and field conditions. Additionally, investigating temporal anomaly detection for disease progression monitoring and extending the approach to multi-spectral or hyperspectral imagery could enhance detection capabilities, particularly for early-stage infections. Finally, developing integrated systems that combine anomaly detection with targeted classification for identified anomalies could create more comprehensive disease management solutions for practical agricultural applications.

In conclusion, this study establishes that computationally efficient feature extraction architectures, when combined with appropriate dimensionality reduction and anomaly detection algorithms, can effectively identify plant diseases across varying imaging conditions. These findings provide a foundation for developing practical, field-deployable systems for early disease detection that can contribute to sustainable agricultural practices and improved food security.

Bibliography

[1] Models and pre-trained weights — Torchvision main documentation.

[2] facebookresearch/dinov2, March 2025. original-date: 2023-03-29T16:00:37Z.

[3] Jayme G. A. Barbedo. Factors influencing the use of deep learning for plant disease recognition. *Biosystems Engineering*, 172:84–91, August 2018.

[4] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May 2000.

[5] Samuele Bumbaca and Enrico Borgogno-Mondino. Supporting Screening of New Plant Protection Products through a Multispectral Photogrammetric Approach Integrated with AI. *Agronomy*, 14(2):306, February 2024. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[6] Raghavendra Chalapathy and Sanjay Chawla. Deep Learning for Anomaly Detection: A Survey, January 2019. arXiv:1901.03407 [cs].

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[8] Konstantinos P. Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, February 2018.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs].

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. pages 4700–4708, 2017.

[11] David P. Hughes and Marcel Salathe. An open access repository of images on plant health to enable the development of mobile disease diagnostics, April 2016. arXiv:1511.08060 [cs].

[12] Ryoya Katafuchi and Terumasa Tokunaga. Image-based Plant Disease Diagnosis with Unsupervised Anomaly Detection Based on Reconstructability of Colors, September 2021. arXiv:2011.14306 [cs].

[13] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. arXiv:2103.14030 [cs].

[15] Federico Martinelli, Riccardo Scalenghe, Salvatore Davino, Stefano Panno, Giuseppe Scuderi, Paolo Ruisi, Paolo Villa, Daniela Stroppiana, Mirco Boschetti, Luiz R. Goulart, Cristina E. Davis, and Abhaya M. Dandekar. Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development*, 35(1):1–25, 2015. Publisher: Springer Verlag/EDP Sciences/INRA.

[16] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science*, 7, September 2016. Publisher: Frontiers.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. arXiv:2304.07193 [cs].

[18] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen,

Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, 109(5):756–795, May 2021. arXiv:2009.11732 [cs].

[19] Serge Savary, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3(3):430–439, March 2019.

[20] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 07 2001.

[21] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. arXiv:1409.1556 [cs].

[22] Aadarsh Kumar Singh, Akhil Rao, Pratik Chattopadhyay, Rahul Maurya, and Lokesh Singh. Effective plant disease diagnosis using Vision Transformer trained with leafy-generative adversarial network-generated images. *Expert Systems with Applications*, 254:124387, November 2024.

[23] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. arXiv:1905.11946 [cs].

[24] Ranjita Thapa, Noah Snavely, Serge Belongie, and Awais

Khan. The Plant Pathology 2020 challenge dataset to classify foliar disease of apples, April 2020. arXiv:2004.11958 [cs].

[25] Yosuke Toda and Fumio Okura. How Convolutional Neural Networks Diagnose Plant Disease. *Plant Phenomics (Washington, D.C.)*, 2019:9237136, 2019.

[26] Sasikala Vallabhajosyula, Venkatramaphanikumar Sistla, and Venkata Krishna Kishore Kolli. A novel hierarchical framework for plant leaf disease detection using residual vision transformer. *Heliyon*, 10(9):e29912, May 2024.

Conclusions

This thesis has systematically investigated the integration of geomatics techniques with geostatistical methods to improve PPP efficacy and selectivity evaluations. Through a series of complementary case studies, we have demonstrated the feasibility and practical requirements for leveraging spatial data to more effectively model and exclude environmental variability from statistical analyses.

The first case study established that photogrammetry combined with machine learning provides a robust methodology for obtaining spatial coordinates alongside continuous variable observations, specifically plant counts. Our research determined precise dataset requirements for deploying these technologies within the EPPO framework. We found that transformer-mixed architectures (RT-DETR) required fewer training samples (approximately 60) to achieve benchmark performance compared to pure CNN models like YOLO variants, which needed substantially more samples (110-130). Importantly, we demonstrated that in-domain training data is essential for reliable performance, as no out-of-distribution trained model achieved the EPPO benchmark regardless of architecture. This

finding has significant implications for the future of digital data collection in phytosanitary trials, as it suggests that even with advanced models, the need for domain-specific training data remains critical for achieving accurate and reliable results.

The second case study (ordinal variables) explored the automation of phytotoxicity scoring using geomatics techniques. The integration of photogrammetric 3D modeling with multispectral imaging enabled accurate reproduction of visual assessments (κ > 0.7) while providing precise spatial coordinates for each observation. This geomatic approach not only automated subjective evaluations but also facilitated the conversion of ordinal data to continuous scales, enabling more powerful parametric statistical analyses within geostatistical frameworks.

The third case study (binary and nominal variables) demonstrated the effectiveness of combining pre-trained machine learning models with georeferenced data for plant health classification. The geomatic workflow provided spatial context for anomaly detection algorithms, achieving accuracy > 0.85 without requiring task-specific training data. This approach particularly benefits from the spatial organization of data that geomatics provides, enabling more robust unsupervised learning strategies.

4.1 Geomatic Contributions and Innovations

This research has demonstrated that geomatics techniques provide three critical advantages for PPP trials:

Spatial Data Integration: Photogrammetry and spectral imaging workflows automatically generate precise spatial coordinates alongside obser-

vations, eliminating the traditional barrier to implementing geostatistical methods in agricultural trials. Our findings show that centimeter-level accuracy is achievable across all EPPO variable types.

Enhanced Data Density: Geomatics techniques enable collection of thousands of spatially-referenced observations compared to traditional manual methods. This dramatic increase in sample size improves statistical power while maintaining spatial independence through appropriate geostatistical modeling.

Reproducibility and Standardization: Digital geomatics workflows provide objective, repeatable measurements that reduce human bias and improve inter-observer consistency. The georeferenced datasets enable retrospective analysis and validation, supporting regulatory requirements for data quality and traceability.

4.2 Future Research Directions

Several promising avenues emerge from this work:

Temporal Geostatistics: Extending spatial modeling to include temporal dimensions using time-series drone surveys could improve understanding of treatment effects over crop development cycles.

Multi-sensor Fusion: Combining thermal, LiDAR, and hyperspectral data within unified geomatic workflows may enable detection of subtle stress responses currently missed by visual assessments.

Unsupervised ML: Global development of self-supervised and weakly-supervised learning techniques could further reduce the need for exten-

sive labeled datasets, enabling more efficient model training across diverse agricultural contexts.

Regulatory Integration: Collaborating with EPPO to develop standardized protocols for digital data validation and acceptance within regulatory frameworks.

In conclusion, this thesis establishes geomatics as a transformative technology for PPP evaluation, providing practical solutions to longstanding challenges in agricultural statistics. The integration of spatial data collection with geostatistical analysis represents a paradigm shift toward more robust, objective, and statistically powerful efficacy assessments that better serve agricultural research and regulatory decision-making.

A special acknowledgment goes to SAGEA, for founding this research and offering me the opportunity to collaborate in such a dynamic and enriching environment. Their support has been invaluable in shaping my journey. Finally, I would like to thank my tutor and colleagues, with whom I’ve shared ideas, challenges, and most importantly, the satisfaction of collective growth.