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

Contents

Chapter 1

Introduction

1.1 Phytosanitary Products

Phytosanitary products, commonly used as a synonym for ”Plant Protection Products” (PPPs), are a specific category of pesticides 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, phytosanitary products 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 phytosanitary products.

Phytosanitary products contain at least one active substance, which can be either chemical compounds or microorganisms, including

viruses, that enable the product 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 phytosanitary products can be distinguished based on the type of organism they target or the function they perform, including:

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

The parameters identified through the risk assessment are compared with the values established by directive 97/57/EC **?**, 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 **?**).

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 **?** recognizes the International Plant Protection Convention (IPPC) as the only international institution in charge of emitting standards for plant health **?**. 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 **?**.

1.2 EPPO Standards

Generics on efficacy assessments are reported in PP 1/181(5) **?**, which describes herbicide, fungicide, bactericide, and insecticide efficacy on the target evaluation. PP 1/135(4) **?** describes the selectivity assessment procedures, in other words: the standard phytotoxi-

city assessments of PPPs. The PP 1/152 **?** 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 thesis (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**

1.2.1 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’” **?**. 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 control and reference thesis. The randomization of thesis 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 phytosanitary product 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 phytosanitary product 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.

1.2.2 Sampling Method and Measures Units

After defining the experimental units through the randomization design choise, the next step is to define the sampling method and the measures units. 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 thesis and control expressed in percentage. As specified by PP 1/152 **?**, 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) **?** "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 **?** and other ones. In general, data types must undergo the classification presented in Table 1.1

Table 1.1: Different modes of observation and types of variables

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

1.2.3 Statistical Analysis

The statistical analysis of trials is equally critical, providing objective assessment of treatment effects. While PP 1/152 **?** 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 ad-

ditivity 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.

Through adherence to these rigorous design and analysis standards, researchers can generate reliable evidence to support phytosanitary product registration while ensuring that products demonstrate consistent efficacy across relevant agricultural conditions.

1.3 Geomatics Techniques

Traditional statistical analysis for phytosanitary product trials still relies on Fisher's principles of experimental design **???**, 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 **??**. The experimental design part that mostly relies on human choice is the block disposal and the control arrangement **??**. The block disposal should guarantee that the environmental variability is minimized within the block and maximized between blocks **??**, while the control arrangement ensures that the untreated control is not influenced by adjacent treated plots **??**. Problems arise when environmental variability effects, unobserved during set-up, make the parametric statistical analysis invalid due to heteroscedasticity of the residuals **??**. Often in such cases, shifting to non-parametric tests could mean a decrease of power **??**. If the experimental design and the statistical model could be able to catch the environmental variability ”a posteriori” instead of ”a priori” **??**, the statistical analysis should be more robust, reliable and free from unexpected heterogeneity whithin the blocks. Variograms are fundamental tools in geostatistics that quantify the spatial dependence of a random field **??**. They characterize how data similarity changes with distance, making possible to include the indipendent from the thesis spatial variation in a parametric model. The empirical (sample) variogram $\hat{\gamma}(h)$ is calculated by: [(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 **??**.

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 **??**. 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 **??**, 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 **??**. 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 **??** 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 **??**. If the only imbricated control is present, area size of control is even more important, because having it smaller than thesis 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 **?????**, 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) **???**. 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) **?**, SpATS has been effectively applied to field trials with varying dimensions, but there are practical considerations for minimum requirements: a minimum grid of 5×5 (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. The bigger drawback of adopting a geostatistic approach to exclude environmental variability is the need for georeferenced observations **??**, often impractical for traditional assessments like visual ones. One can argue that for trials with regular grid layouts on flat fields, the geographic coordinates of the plots are not needed, as the distance between plots is regular and can be taken as unitary to get spatial coordinates. However, geostatistic models are sensitive to accuracy and precision of coordinates, so they must be scaled to the environmental feature minimum spatial size. Since geostatistical approaches are adopted precisely to address unknown patterns of environmental variability, we cannot set a minimum coordinates precision ”a priori”. Today, digital technologies such as georeferenced imaging, ensure a very high precision of measurments **?**. Digital approaches can automate data collection and analysis, improving the reproducibility of results, ultimately accelerating the development and registration of effective phytosanitary products. While the EPPO experimental design standards provide a solid foundation for

conducting phytosanitary product trials, the increasing availability of digital technologies offers new opportunities to enhance the quality (in the ”Quality of a mode of observation” sense **?**) and efficiency of these assessments.

Another advantage of adopting digital approaches is the dramatic increase in the number of observations that can be collected. Unlike manual methods, which are inherently limited by human capacity and time constraints, automated and semi-automated systems can continuously gather data with minimal interruption. This greater volume of data not only improves the resolution and granularity of analysis but also significantly enhances the statistical power of hypothesis testing.

The power of a statistical test, defined as the probability of correctly rejecting the null hypothesis when it is false, directly depends on the sample size (number of observations), as noted by a classical statisticians as Fisher **?**. However effiecient technics to collect data does not garantee indipendency of samples. In other words, having a powerfull tool to collect data does not mean that we can rely on a higher amount of replications.

Wheter repeated observations per experimental unit (pseudo-replications) are produced, Generalized Linear Mixed Models (GLMM) **??** should be used to benefit from digital observations size enhancement. GLMMs are a powerful extension of GLMs that can account for the correlation structure of repeated observation by incorporating random effects, thus providing more accurate estimates of treatment effects and their associated uncertainty. All the described geostatistical techniques

can be seamlessly integrated with GLMMs to provide a comprehensive analysis of spatial-temporal data, enhancing the accuracy and reliability of treatment effect estimates.

To regulate the use of this kind of technologies, the EPPO published a new standard, PP 1/333(1) **?**, which filled the gap in the use of digital technologies in phytosanitary product 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**: Coefficient of determination (R^2) higher than 0.85.
- **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. Perform metric measurements in agriculture is probability the anciest problem that human faced with geometry and the rise of geography. Photogrammetry is a geomatic technique that allows the acquisition of spatial data in metric scale by processing and analyzing georeferenced photographic images. It allows to create digital maps of agriculture landscapes. This kinds of maps that will be better depicted in the following sections, can be analyzed to track and measure the spatial distribution of the variables of interest. The most effective technics to do so are Machine Learning (ML) and Geostatistics. ML is a branch of artificial intelligence that allows counting and classifying. Through ML, it is possible to implemtent digital tools that can boost the data collection. However, adoption of digital tools is not only a matter of data collection but also of data analysis. In many case, perform classical statistical analysis on digital data with reduce the benefits of collecting spatial data or simply is not possible. Geostatistics, as already shown, offers a wide range of statistical methods that can be used to overcome the limitations of traditional statistical methods and to exploit the full potential of digital data. In the following sections, we will explore the opportunities and con-

straints of deploying geomatic techniques to increase the efficacy of phytosanitary products.

1.3.1 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 **??**.

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 citehartley$_{m}ultiple_{2}003, szeliski_{c}omputer_{2}010.$

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 **??**.

Image orientation involves determining the position and orientation of the camera at the time each photograph was taken. This pro-

cess is divided into two main steps: interior orientation and exterior orientation **??**.

- **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. This is achieved by identifying and matching common points (tie points) in overlapping images and using these points to solve for the camera parameters.

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 **??**.

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 **??**:

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

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 whithin images was performed manually. Today many algorithms doing 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) **???** 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) **??**. 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 **??**. 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.

1.3.2 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 **??**. Multispectral images are typically acquired using specialized cameras or sensors that can capture light in different spectral bands, ranging from ultraviolet (UV) to infrared (IR) wavelengths **??**. Some study tested also the feasibility to use bands in the termal infrared (TIR) range **?**. 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 material 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 **?**. 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 near-infrared (NIR) bands of the multispectral image **??**. 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 values 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), Soil-Adjusted Vegetation Index (SAVI), and Leaf Area Index (LAI), each providing specific information about vegetation health and condition **??**. These indices can be used to monitor crop performance, assess nutrient status, and evaluate the impact of environmental factors on plant growth **??**.

1.3.3 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 **?**. Unlike traditional programming, where explicit instructions dictate the output for given inputs, ML models identify patterns and relationships within data to generate predictive outcomes **?**. 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 **?** and plant protection **?**. Within the context of phytosanitary product 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 **?**. 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 **???**.

The primary objective of employing ML techniques in phytosanitary product trials is to improve accuracy, precision, and reproducibility while reducing manual intervention and subjective bias **?**. 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 **?**. For example, generalized linear models (GLMs) can be enhanced with ML techniques to improve their accuracy and adaptability **?**. Deep Learning (DL) is a subfield of ML that studies a particular class of models named Deep Neural Networks **?**, 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 phytosanitary product efficacy evaluation, computer vision methods are increasingly used for automated observation and measurement, particularly when integrated with digital imaging and photogrammetry **?**. The use of computer vision within PP1/333 trials significantly enhances data acquisition by enabling digital sensing and precise measurement of crop conditions **??**. Techniques such as image segmentation, object detection, and texture analysis can automatically identify plant stress, disease symptoms, and pest damage **?**.

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 **?**.

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 plant protection products and phytopathometry research, struggle to compile sufficiently large and diverse training datasets **?**. 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 **?**. 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 **?**. These models can effectively perform new tasks with minimal fine-tuning, a phenomenon known as ”few-shot learning” or ”in-context learning” **?**. 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 **?**. Self-supervised learning, while promising to revolutionize machine learning by eliminating the need for manual labeling, presents its own set of challenges **?**. The computational resources required for training large self-supervised models are substantial, often exceeding the capabilities of smaller research labs or specialized studies **?**. This computational intensity creates a significant barrier to entry, particularly for domain-specific research like phytosanitary product efficacy evaluation, where the computational and expertise requirements may outstrip the available resources of a typical research group **?**. 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 **?**. Researchers can exploit the rich representations learned by foundation models, applying them as powerful feature extraction mechanisms across various downstream tasks **?**. 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 **?**. This transfer learning approach allows for efficient model adaptation, reducing the need for extensive domain-specific data collection and annotation **?**. In the context of specialized fields like phytosan-

itary 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 **?**. 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 **?**. 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 **?**.

1.4 The Literature Gap and the Thesis Aims

Despite the clear benefits of integrating geomatics techniques into phytosanitary product efficacy trials, the high amount of observations needed for The aim of this thesis is to investigate the limits of integrating geomatics techniques in the design and analysis of phytosanitary product efficacy trials. As already discussed, the EPPO standards provide a solid foundation for conducting experimental trials, but the increasing availability of digital tools and technologies offers new opportunities to enhance the quality and efficiency of these assessments. By leveraging geomatics techniques such as photogrammetry, geostatistics, and machine learning, researchers can improve data collection, analysis, and interpretation, ultimately ac-

celerating the development and registration of effective phytosanitary products. Throughout a study case each variable type, we will explore the opportunities and constraints of deploy geomatic techics for increase phytosanitary products effects estimation.

Bibliography

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.

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. Pages: 1–32 Volume: L 230.

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

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

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

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.

European Commission. Uniform Principles for evaluation and authorisation of plant protection products, 1997. Pages: 87–109 Volume: L 265.

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. Pages: 1–50 Volume: L 309.

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

World Trade Organization. The WTO agreement on the application of sanitary and phytosanitary measures (SPS agreement). *World Trade Organization*, 1995.

Chapter 2

Study Cases

2.1 Continuous Variables

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 The Problem of Plant Counting

Plant counting is a critical operation in precision agriculture, plant breeding, and agronomical evaluation. Accurate plant counts can provide valuable information for both farmers and researchers. This task was often performed manually by human operators. Today, this process can be automated by the use of computer vision algorithms. To validate a method for counting, it is critical to set a benchmark for accuracy. A benchmark can be defined by the accuracy of manual counting, international standards, or by comparison with other already accepted methods. Accuracy of plant manual counting depends on human performance, so variable, but often taken as golden sample. According to the European Plant Protection Organization, the benchmark for acceptance is a coefficient of determination (R^2) of 0.85 when compared to manual counting [8]. This corresponds to a Root Mean Square Error ($RMSE$) of approximately 0.39. Also scientific literature mention a R^2 value of 0.85 with ground truth (manual counting) as a benchmark for acceptance [85].

Literature shows that benchmark for acceptance can be achieved with computer vision object detection [57, 22, 17] or regression models [56, 2]. A superiority of object detection over regression models in terms of accuracy was found [85]. Object detection is also more versatile than regression models, because, object detection model inference on georeferenced orthomosaics delivers plants ge-

ographical coordinates, not only density per area. The validation of this ability rely on metrics as Intersection over Union (IoU), Average Precision (AP), and Average Recall [54] rather than the coefficient of determination. Identification supports are usually bounding boxes, which are rectangles that enclose the object of interest, but they can also be points or other kind of geometries. Counting of plants is then commonly performed by counting the number of geometries that enclose the objects of interest (the plant) in a image area. Georeferenced orthomosaics are images created through aerial photogrammetry, a process that involves capturing overlapping georeferenced images taken by nadiral view picturing georeferenced ground control points, and performing bundle adjustment to form a single, seamless image [46]. Georeferenced orthomosaics are scaled and oriented to a geographical coordinate system. It implies that the pixel coordinates corresponds to geographical coordinates and are projectables in a metric system. Using georeferenced orthomosaics in seedling counting makes object detection easier, because the metric scale can be used to locate the objects in the image and prospective variance is reduced to zero.

2.1.1.2 Case study: Maize Seedling Counting

Plant counting is affected, as many object detection application fields, from data scarcity: public datasets are rare and often not suitable for the task because of the large environmental variability and their images are not orthorectified or scale is unknown. Even so, some useful dataset for training a plant counting object detector can be

found in public repositories [32]. These dataset may come as a part of a scientific study or with poor technical specifications. Selecting a specific crop at a specific growth stage can reduce the variability that the model should learn and make it possible to better study the other variables that can affect the dataset size and quality requirements. For this study we choose grain maize seedlings (Zea mays L.) at the V3-V5 (BBCH 13-15) growth stage [61], because it is the most represented plant in scientific [22, 57] and not scientific [5, 4] open datasets from aerial photogrammetry. Maize is a commodity crop that is widely grown in the world and is the most important crop in the world by production [25]. Grain maize seedlings in that stage are easy to count because of their low overlapping and fixed intra-row and inter-row spacing, differently by silo-maize seedling that is seeded with extremely low inter-row spacing. This particular seedling configuration that makes grain maize a good candidate for object detection, is shared with other crops that are seeded in rows with inter-row spacing such as sunflower (Helianthus annuus L.) or sugarbeet (Beta vulgaris L.).

2.1.1.3 Object Detection approaches

The development of object detection algorithms has evolved from not machine learning methods, here named handcrafted methods (HC), to modern deep learning-based (DL) techniques. Today, all the state-of-the-art object detection methods are based on DL models. Nevertheless, HC methods are still used in some cases [22, 29]. Most of modern DL object detection uses convolutional neural net-

works (CNN) [48] based frameworks (e.g., Faster R-CNN [3], YOLO [7]) or Transformer [72] based approaches (e.g., DETR [19]) or mixtures of both approaches. The main difference between CNN-based and Transformer-based models is the way they process the image. CNN-based models process the image in a grid-like fashion (convolutions), while Transformer-based models process the image as a sequence of patches (attention mechanisms) [23]. On common benchmarks as COCO [54], PASCAL VOC [38], and ImageNet [1], Transformer-based models have shown to be more accurate than CNN-based models [84] . CNN-based models are still widely used in object detection, because they are more efficient in processing small images and have a lower computational cost [43]. However, when fine-tuning with scarce data, Transformer-based object detectors generally perform better than CNN-based detectors, provided they are pretrained on large datasets [66, 51, 12].

To represent the categories and compare performance between pure-CNN and Transformer-mixed architectures that are effectively used as object detectors in real plant counting applications, YOLOv5 and YOLOv8 can be taken as pure-CNN architecture representatives for their large use in agriculture [15]. Their large diffusion in agriculture applications is justified by the fact that they leverage good precision and the low need in terms of dataset size in respect other CNN architectures [71, 53, 81]. Real-Time-DETR (RT-DETR) is a recent Transformer-mixed architecture that outperforms YOLOv5 and YOLOv8 [83]. YOLO11 has been recently proposed as a Transformer-mixed YOLO architecture that outperforms RT-DETR on COCO dataset [44].

Recently, zero-shot and few-shot object detection have emerged as promising paradigms to alleviate the need for large annotated datasets. While traditional object detection models (many-shots object detectors) require extensive labeled data for training, few-shot and zero-shot object detection aim to detect novel objects with little to no labeled examples respectively. The term ”shot” refers to the number of annotations used to train the model over all the images. Each shot corresponds to an individual, that in the case of few-shots, is used to prototype the object of interest. Zero and few-shots approaches often leverage feature transfer or meta-learning components to generalize across classes under extreme data scarcity [52]. This approach reduces the annotation burden and is especially beneficial in domains where collecting exhaustive training data is impractical. Zero-shot object detection detects new categories without any training samples by leveraging semantic relationships or contextual information learned from known classes [16]. Few-shot approaches optimize models to learn quickly from a handful of labeled examples [35]. Meta learning is a common approach in few-shot object detection [6, 78, 27, 80], while Cross-Domain Few-Shot Object Detection (CD-FSOD) has recently surpassed this approach by leveraging domain adaptation techniques [68]. DE-ViT [79] and CD-ViTO [28] are the latest models that have shown promising results in few-shot object detection. Zero-shot actual state-of-the-art object detector models are YOLO-World [40] , OWLv2 [62] , and Grounding DINO [55]. OWLv2 (Open-World Localization v2) is an zero-shot object detector that represented a significant evolution in open-vocabulary detection capabilities.

2.1.1.4 Dataset Size and Quality Requirements

As already mentioned, the performance comparison between object detection models is usually based on differences in metrics such as AP, calculated on standard datasets that are not representative of the agricultural application field. Many studies have been conducted on object detection for plants in open field [17, 30, 37, 42, 45, 49, 50, 57, 56, 58, 59, 73, 74, 82], but few have argued or focused on dataset minimum requirements for training a robust plant object detector [22, 13]. Some study focused on few-shots approach for plant counting [74, 11], but only two specifically accounting on few-shots method for maize seedling counting [41, 75]. Unfortunately, the lonely two studies evaluating few-shot performance on maize seedling counting by orthomosaics do not achieve the benchmark and do not clearly specify the number of shots used. Also no zero-shot benchmark for maize seedling counting has been set yet, and no research on this application has been done yet.

This critical research gap presents significant challenges for practitioners in precision agriculture who must decide how much data to collect and annotate for effective maize seedling detection. Without systematic evaluation of minimum dataset requirements across different detection approaches (many-shot, few-shot, and zero-shot), it remains unclear whether resource-intensive manual annotation can be reduced or eliminated. Furthermore, the agricultural domain's unique characteristics: variable environment, and plant phenotype, may fundamentally alter the data requirements compared to general computer vision benchmarks. Determining these requirements

would provide practical guidance for implementing object detection in agricultural workflows while optimizing the trade-off between annotation effort and detection performance.

Even if no research has been made in order to minimize or set a benchmark for the dataset size and quality required to train a maize seedling object detector, it is well known that the performance of a DL model is directly related to the amount of data used for training [70] and its quality [10]. Dataset size and quality predictability in DL has been proven to be addressable with empirical approaches [34, 60]. As already mentioned, model architecture is critical in dataset requirements as different models may require different dataset sizes and qualities to achieve the same performance [64, 18]. Backbone is pivotal on downstream tasks as object detection [24]. The importance of using a domain specific backbone has been proven also for plant/leaves segmentation on orthomosaics [67], but backbone training is still prohibitive for many applications. No backbone weights specialized on agricultural orthomosaics has been published yet, as for the most of the specialized domains. Even if a so specialized backbone exists, some concern will came out about its out-of-distribution generalization capability [33]. So, because of limited resources, today only the use of a general backbone is possible in practical sense [36], even if a decreasing in dataset size requirements is expected with a domain specific backbone because of the out-of-distribution generalization [31]. Another factor that can affect the dataset size and quality requirements other than the already cited model architecture and use of pre-trained backbones, is the training data augmentation strategy. Some studies proposed high-

tly computationally expensive data augmentation such as trainable data augmentation [21] or data augmentation with generative adversarial networks [14]. Other studies proposed to use less computationally expensive data augmentation strategies [69, 20] and someone even proved that random image augmentation can provide equivalent results to more expensive technics [63]. As also image augmentation can lead to prohibitive computational costs, the use of less computationally expensive data augmentation strategies is recommended for fine-tuning to downstream tasks [69].

Nevertheless, the most important factor affecting DL training is the dataset source [70]. It has been observed that using training samples from the same dataset as the inference (in-domain dataset) dramatically increases accuracy and reduce the need for a large dataset in respect to collecting training samples from other datasets (out-of-distribution dataset) [22, 13]. From comparison of the studies here mentioned, dataset source seems to be the most important factor in determining the dataset size and quality needed to achieve the benchmark for acceptance.

2.1.1.5 Study Aim

This study aims to determine the minimum dataset size and quality required to train a object detection model for identifying maize seedlings in georeferenced orthomosaics achieving the benchmarks set by international organizations and recognized in scientific literature. Here, the dataset size and quality are respectively defined as the amount of annotated images in the training set, and the accu-

racy of the annotations. Also the effect of training dataset source will be evaluated in this study. Models architecture and size will be taken into account, as long as the object detection downstream task is concerned. After setting the dataset size and quality minimum requirements for many-shots object detectors, we will evaluate if new zero-shot and few-shot object detection models can achieve the same benchmark for acceptance with less data and annotations. We will also discuss the need for HC method in a DL object detection pipeline.

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 seedlings 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 [61].

This study uses two dataset sources as training sets: the out-of-distribution dataset (OOD) and the in-domain dataset (ID). 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 [56, 22] and from internet repositories [5, 4]. The ID datasets were collected during this study. This ID dataset

creation consisted in 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 @ 10 m AGL (above the ground level). The number of images captured depends on the study area size that was about 2 hectares for ID_1 location and about 1 ha for the other two. For each location, an orthomosaic was created using a photogrammetric software. Bundle adjustment error was estimated as 38 mm using the ground control points surveyed by GNSS operating in VRS-NRTK mode. The orthomosaics were generated with an average ground sampling distance of 5 mm/pixel in the WGS84/UTM 32 N reference system.

The OOD scientific datasets consist of tiles of georeferenced orthomosaics of maize seedlings from 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 annotation 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 on a Geographical Information System (GIS) environment, with the tiles grid overlapping the orthomosaic to focus on target tiles without losing the surrounding context, so without losing bordering plants. An-

notations were created as squared bounding boxes of size length equal to the minimum distance between two plants in the row, with each box centered on an individual seedling stem.

To make the two kind of dataset comparables 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*224 pixels tiles. This tile size was selected because at 5 mm/pixel resolution it covers 1.12×1.12 meters of field area. Given that typical grain maize inter-row distance is 0.75 meters, 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. This particular image size was also chosen as a standard from AlexNet [47] 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 testing dataset, while other 150 tiles are used as training dataset.

2.1.2.2 Handcrafted object detector

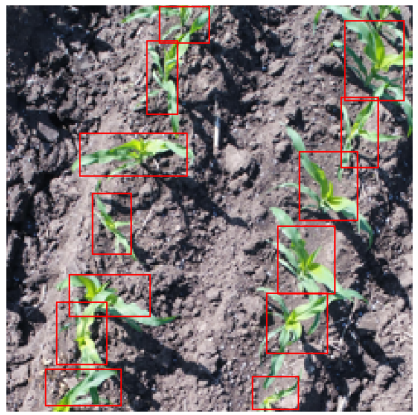
Like other works [22, 29, 56] we wrote an HC algorithm to get annotated tiles from the orthomosaics, basing it on agronomical knowledge and color thresholding. Hue, saturation and value (HSV) color space was used here to threshold the image, to get green pixels, but other color spaces can be used. For the execution of this algorithm, the following graphical and agronomical parameters must be

Table 1: Summary of Datasets Used in the Study

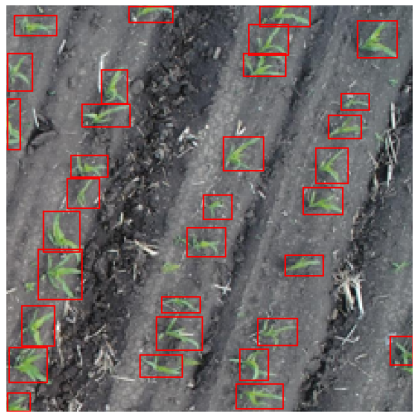
Dataset	Phenological Stage	Train Size	Test Size
OOD Scientific			
DavidEtAl.2021 [22]	V3	182 tiles	N/A
LiuEtAl.2022 [56]	V3	596 tiles	N/A
OOD Internet			
OOD_int_1 [5]	V3	216 tiles	N/A
OOD_int_2 [4]	V5	174 tiles	N/A
ID			
ID_1	V3	150 tiles	20 tiles
ID_2	V3	150 tiles	20 tiles
ID_3	V5	150 tiles	20 tiles

(a) DavidEtAl.2021

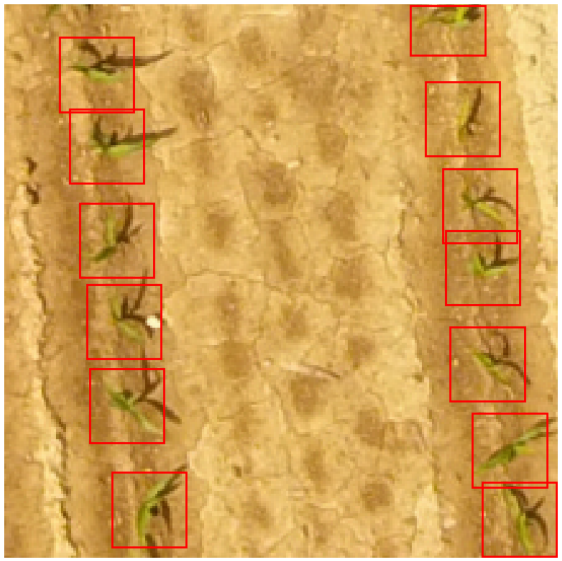
(b) LiuEtAl.2022



(c) Internet Maize stage V3

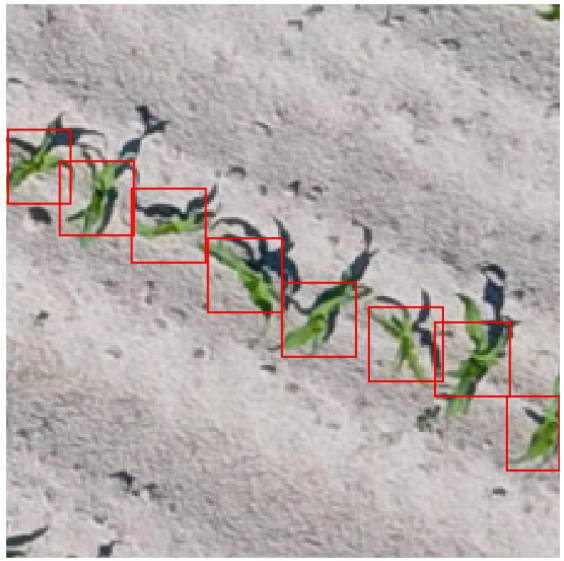


(d) Internet Maize stage V5



(e) ID_1

(f) ID_2



(g) ID_3

Figure 1: Sample images from each dataset used in the study. The top row shows out-of-distribution datasets: **(a)** DavidEtAl.2021, **(b)** LiuEtAl.2022, **(c)** Internet Maize stage V3, and **(d)** Internet Maize stage V5. The bottom row shows in-domain datasets: **(e)** ID_1, **(f)** ID_2, and **(g)** ID_3.

set: color minimum and maximum thresholds (color threshold), the minimum and maximum leaf area for 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 works: on orthomosaics of maize seedlings at the V3-V5 growth stage, with low weeds infestation, with rows having roughly the same angle with meridian and distance between them.

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 **??** 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 to identify 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 automated extraction of high-confidence annotations from the orthomosaics.

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($binary\_image$) [77]
    return $regions$
for $tile$ in $tiles$ **do**
    $mask \leftarrow \{p \in tile \mid color(p) \in col\_range\}$
    $regions \leftarrow connected\_components(mask)$
    $plants \leftarrow \{region \mid region \in regions \wedge region.area \in leaf\_area\_range\}$
return $plants$

2.1.2.3 Deep Learning object detectors

We chose them for the considerations made in Introduction section 2.1.1.3. We used here the Ultralytics implementation of these models because the implementation is open-source [39], and it makes it possible to tune parameters size coherently across all tested architectures.

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

For all the many-shots 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 uti-

Algorithm 2 H2: Part I - Row Detection

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($points, min\_dist$) [26]
    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 3 H2: Part II - Row Verification

```
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
```

lization 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 (±10%), random translation (±10%), random horizontal flip (probability 0.5), HSV color space augmentation (hue ±0.015, saturation ±0.7, value ±0.4), and mosaic augmentation. These augmentations were selected to reflect potential variations in field conditions without introducing unrealistic distortions.

The training dataset was composed of the OOD or ID training dataset tiles. For the dataset size testing, all the annotations were used, while for the dataset quality testing a percentage of the annotations per image was selected and used.

To test the few-shot approach we trained CD-ViTO with multiple model sizes. The size of this model is determined by the backbone used, which can be ViT-S, ViT-B, or ViT-L [65]. We used the implementation of CD-ViTO provided by the authors [28]. In the context of this study a 'shot' correspond to an image with a single annotated plant. We used 1, 5, 10, 30, and 50 shots to train the model. The

shots were randomly selected from the ID manually labeled dataset, then a random annotation was selected from the same image to be used as prototype. All the combinations of shots and ViT backbone were tested on the ID test dataset tiles.

For testing the zero-shot approach we used OWLv2. We took this architecture as zero-shot object detector example as it is the most stable state-of-the-art model for this task [62, 55]. For test OWLv2 we used the implementation of the transformer library [76] with the parameters published by the authors. We tested the encoder sizes ViT-B/16, ViT-L/14 with the following three pre-training strategies:

- **Base models**: Trained using self-supervised learning with the OWL-ST method, which generates 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 the 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 the following:

- ”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”

All the combinations (here named model settings) of encoder size, pre-training strategy, and text prompt were tested on the ID test dataset tiles.

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

Table 2: Summary of Tested Architectures and Model Sizes[1]

Architecture	Shots	n[2]	s[2] or S	m[2] or B	l[2] or L	x[2]
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[3]	86.0[4]	307.0[5]	-
OWLv2	zero	-	-	86.0[4]	307.0[5]	-

[1] Values represent millions of parameters

[2] Model size variants stand for nano (n), small (s), medium (m), large (l), and extra-large (x)

[3] ViT-S (Small) backbone

[4] ViT-B (Base) backbone

[5] 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 seedlings counting, we conducted a series of experiments where the above mentioned DL models were recursivelly fitted with increasing dataset size and quality. For many-shots models we consider a training dataset split of 10% validation and 90% training, while for few-shots the number of shots determined the amount of training samples. Zero-shots relied only on descriptions in natural language of the objects to be detected. For what concerns only the dataset size eval-

uation, for many-shots models we considered sizes from 10 to 150 images in 15 steps of 10 images, while for few-shots models we considered 1, 5, 10, 30, and 50 shots. For what 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.

For all the models we evaluated the relationship between dataset size or quality and model performance using R^2 and mAP, respectively for plant countig and plant detection. Wheter R^2 provided values below -1 we also considered $RMSE$ as metric for counting. MAPE was considered for few-shots and zero-shots 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.

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

where $RMSE$ measures the average magnitude of prediction errors in the original units (number of plants).

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

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 lower percentage of false positive or false negative. 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.

For object detection performance, we used the standard COCO evaluation metric:

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

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 seedlings 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 [60]. For clarity we list here the functions tested:

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

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

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

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 [9]. The SAHI method slices the testing image into smaller overlapping segments (patches) of the same size as the training tiles and then tests the model on each of them. The model outputs from each patch are then merged by non-maximum suppression and cropped by the original tile extension. The use of such a method is justified by the fact that the model is trained on tiles and the testing dataset is composed by tiles, but the real application is on orthomosaics, so the same object can be present in more than one tile in a cutted (and occluded) way. The SAHI method overcomes this problem ensuring all the possible objects are evaluated by the model as a whole. Thus it is expected to give a better performance in respect to the use of the single tile as input for the model. The prediction were then thresholded by a list of confidence score thresholds to get 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 whithin the thresholds was considered as the model performance for that experiment.

2.1.3 Results

2.1.3.1 Handcrafted object detector

To evaluate the HC object detector as a training dataset extractor from the ID dataset, we measure the amount of annotated tiles that the HC algorithm can extract from the orthomosaics and the accuracy it delivers in comparison to handmade annotations. 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%

Overall the HC object detector performed well on the ID datasets, 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. The HC algorithm was able to extract a significant amount of annotated tiles from the orthomosaics, with a percentage of the dataset ranging from 1.8% to 7.8%. In nominal scale the number of tiles successfully annotated

by the HC algorithm was not constant, but always over 150 tiles, so we took this minimum amount as maximum dataset size for the many-shots training.

2.1.3.2 Many-shots 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 determinations 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, neither the increasing dataset size seems 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 benchmark.

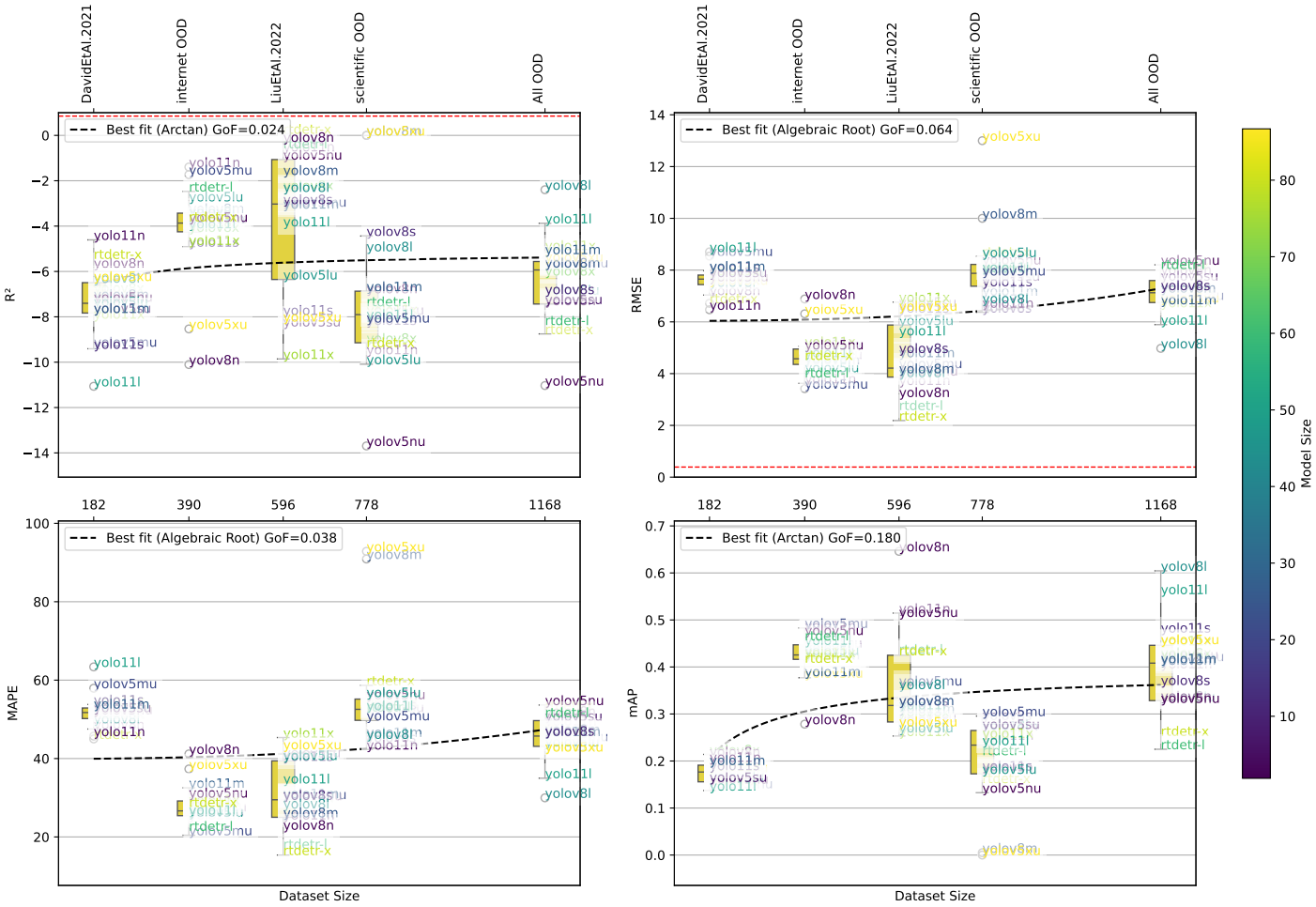


Figure 2: Performance of the many-shots object detection models trained on the different out-of-distribution (OOD) datasets. Subplots represent: R², $RMSE$, MAPE, and mAP respectively at the right top, left top, left bottom, and right bottom. Each subplot contains the boxplots positioned at the corresponding dataset size values and indicating the distribution of all the models prediction metric values for each dataset. Each data point is annotated with the , colored according to the model size. Benchmark thresholds are indicated with red dashed horizontal lines for R² (0.85) and $RMSE$ (0.39). Best fit lines for each metric are plotted using different fitting functions (logarithmic, arctan, and algebraic root), 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, 4, 5 and 6. 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 whitin 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 7. 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.

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

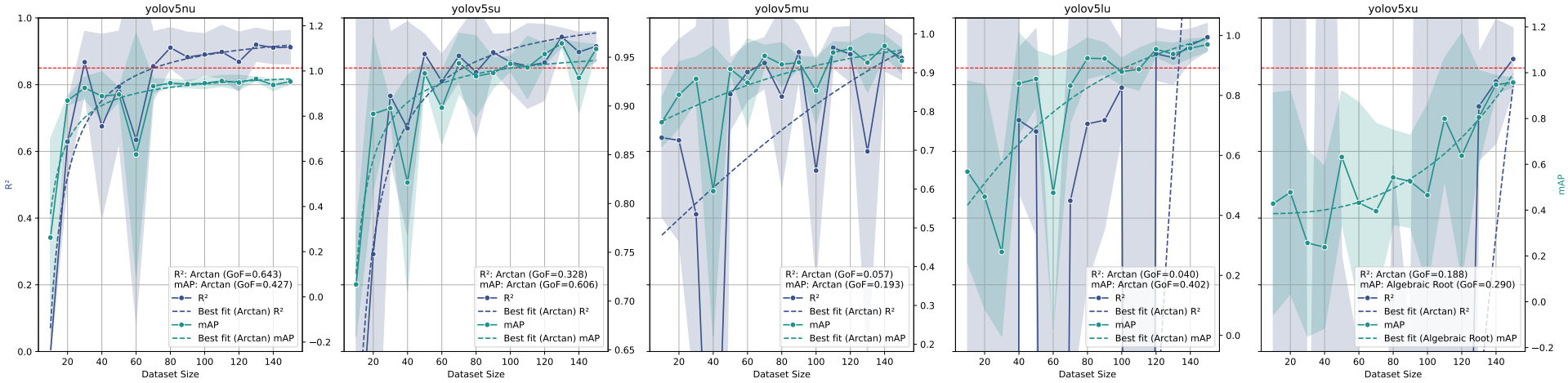


Figure 3: Relationship between dataset size and model performance for YOLOv5 trained and tested on ID datasets. Each subplot represents a different parameters size of the model, increasing from the left to the right. The x-axis represents the dataset size, while the left and right y-axis represents the 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 legend shows the goodness of fit (GoF) for both R^2 and mAP.

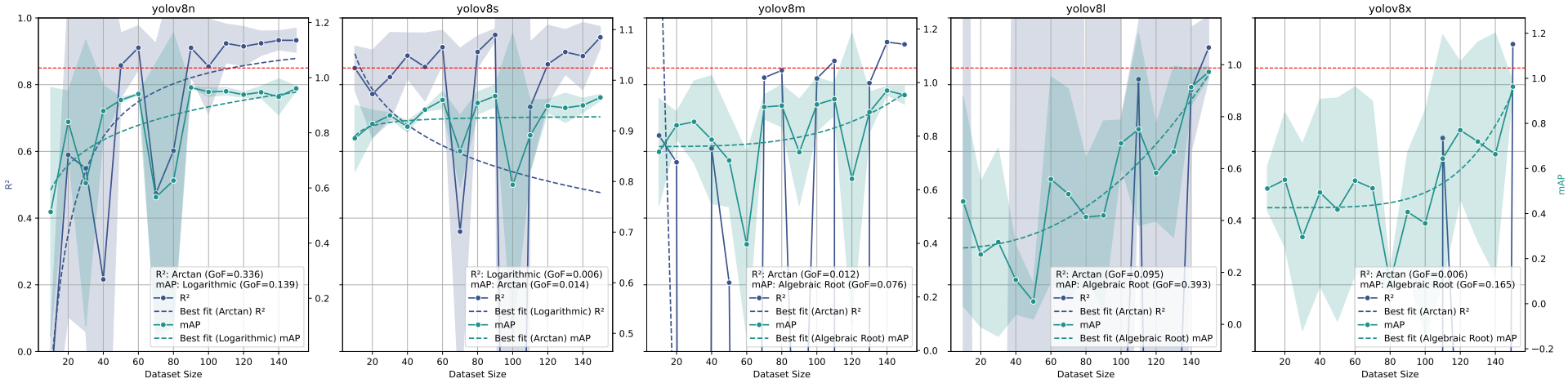


Figure 4: Relationship between dataset size and model performance for YOLOv8 trained and tested on ID datasets. Each subplot represents a different parameters size of the model, increasing from the left to the right. The x-axis represents the dataset size, while the left and right y-axis represents the 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 legend shows the goodness of fit (GoF) for both R^2 and mAP.

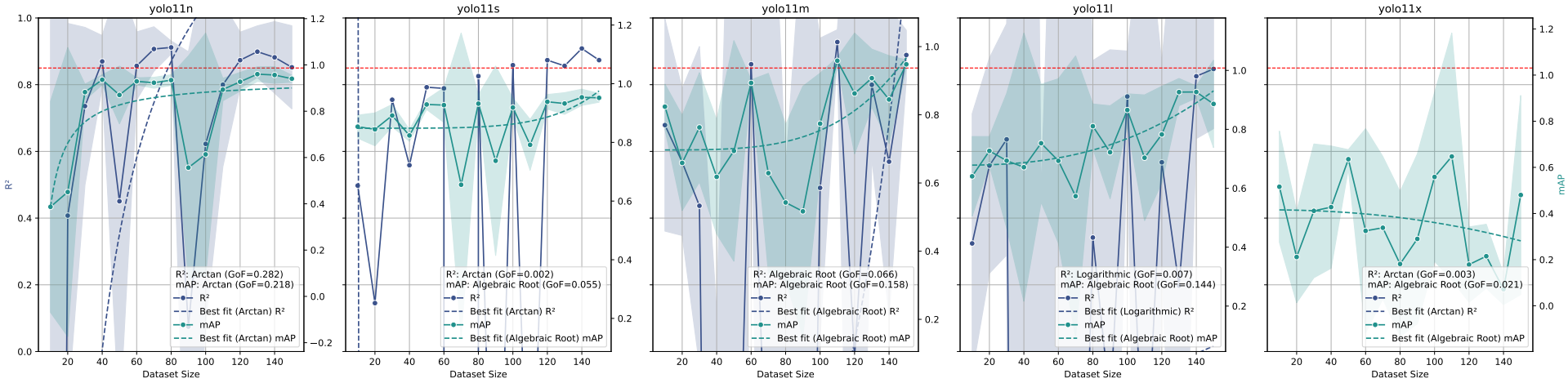


Figure 5: Relationship between dataset size and model performance for YOLO11 trained and tested on ID datasets. Each subplot represents a different parameters size of the model, increasing from the left to the right. The x-axis represents the dataset size, while the left and right y-axis represents the 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 legend shows the goodness of fit (GoF) for both R^2 and mAP.

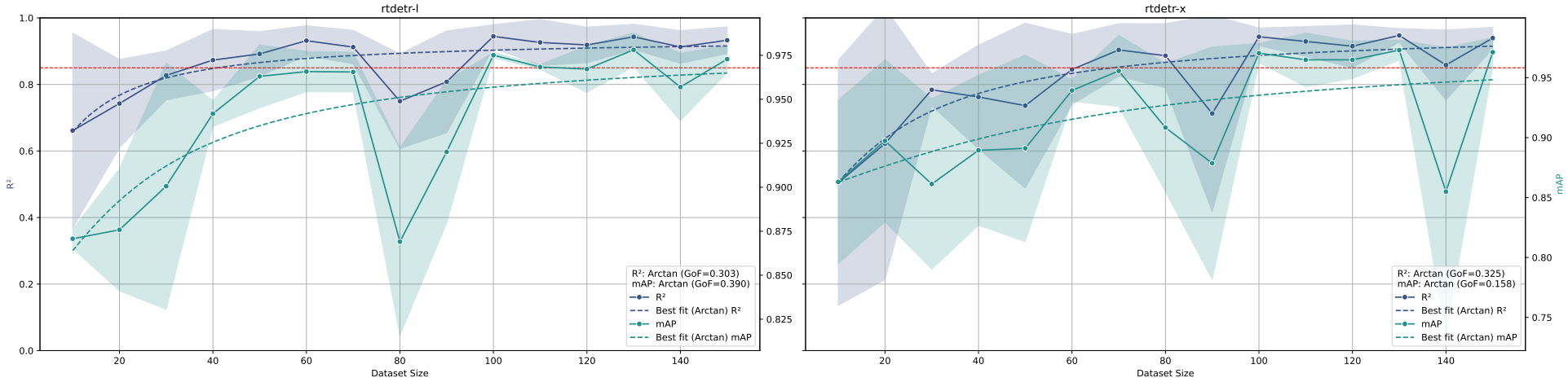


Figure 6: Relationship between dataset size and model performance for RT-DETR trained and tested on ID datasets. Each subplot represents a different parameters size of the model, increasing from the left to the right. The x-axis represents the dataset size, while the left and right y-axis represents 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 legend shows the goodness of fit (GoF) for both R^2 and mAP.

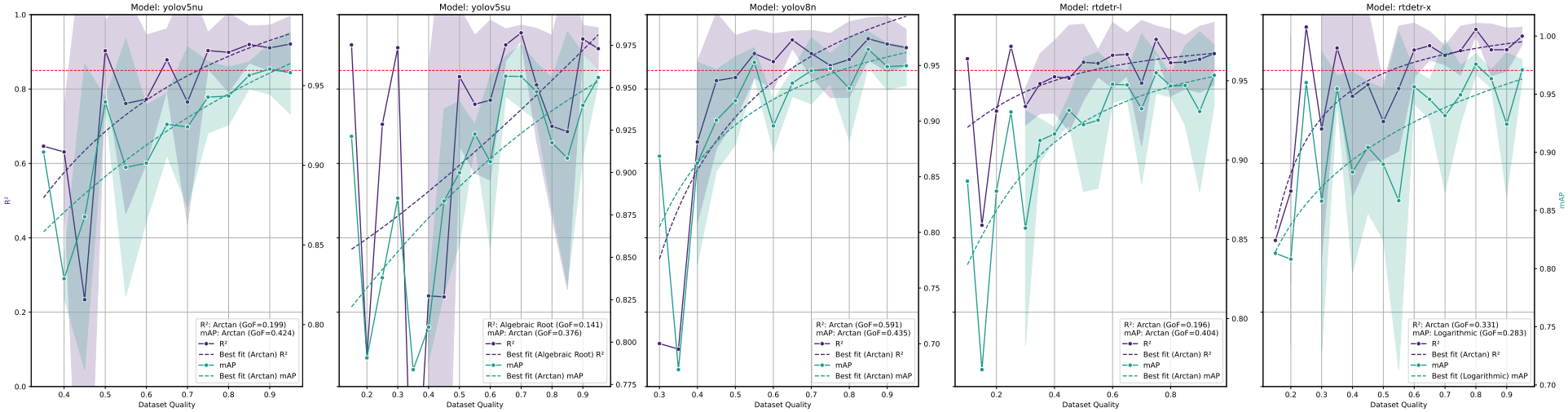


Figure 7: Relationship between dataset quality and model performance for all object detection models that achieved the benchmark. The x-axis represents the dataset quality, while the left y-axis represents the R^2 values. The red dashed horizontal line represents the benchmark R^2 value of 0.85. The legend in the lower right corner of the subplot shows the goodness of fit (GoF) for R^2.

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 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 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.

2.1.3.3 Few-shots object detectors

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

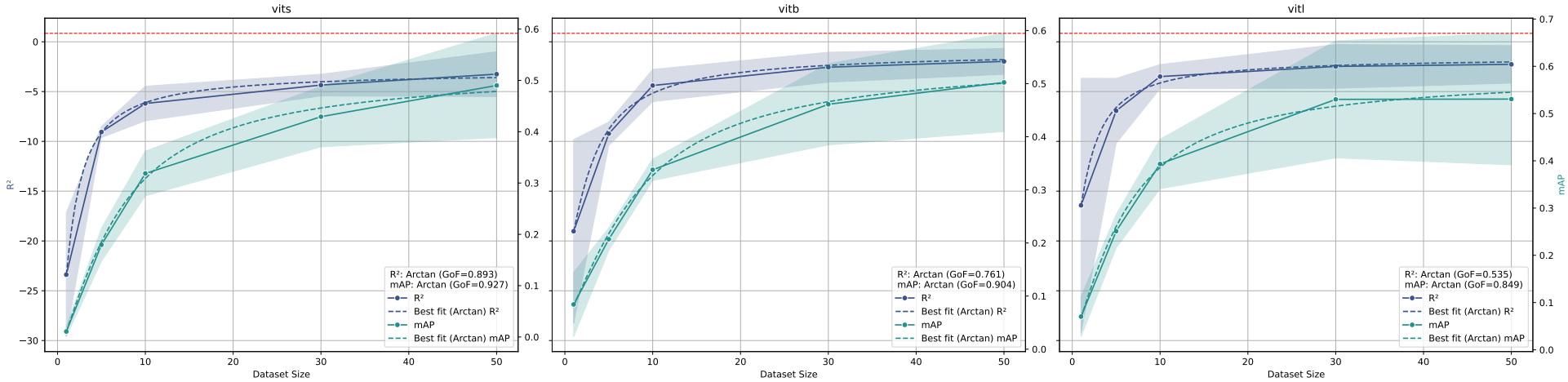


Figure 8: The figure shows the relationship between shots and model performance for the CD-ViTO model trained and tested on ID datasets. The x-axis represents the number of shots. The solid lines represent the mean values, while the dashed lines indicate the shots amount/metric prediction model. The 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.

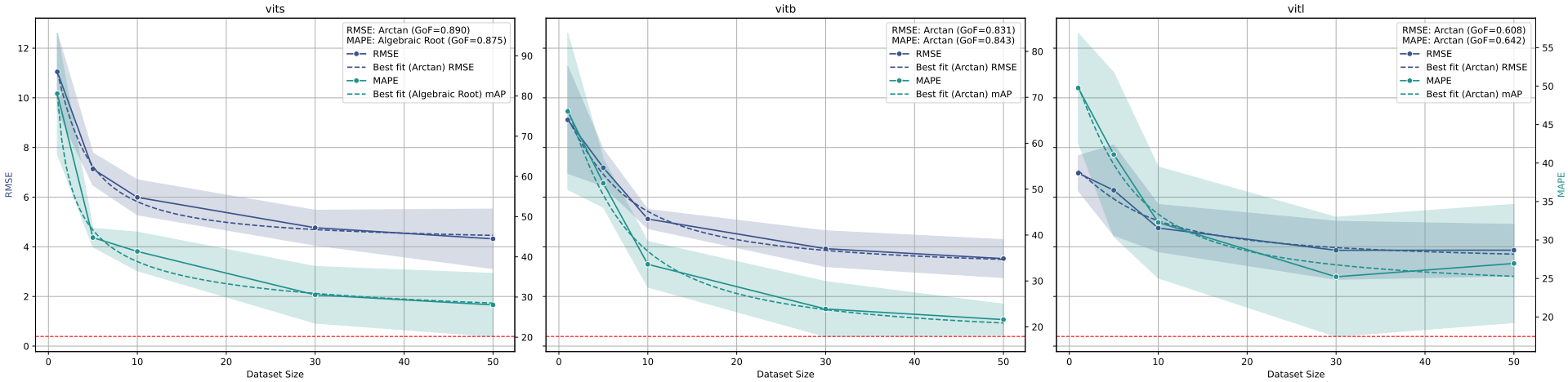


Figure 9: The figure shows the relationship between shots and model performance for the CD-ViTO model trained and tested on ID datasets. The x-axis represents the number of shots. The solid lines represent the mean values, while the dashed lines indicate the shots amount/metric prediction model. The 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.

The best result achieved by the CD-ViTO model was a $RMSE$ 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 25and a mAP of about 0.5 (Figures 8 and 9). It corresponds roughly to a miscounted plant over four as it is visible looking to some predictions of this model in Figure 10. 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-shots object detectors

Figure 11 shows the relationship between the zero-shots 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 as zero-shots models do not require fine-tuning training data. None of the zero-shots 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-shots and few-shots models. Additionally, MAPE values were

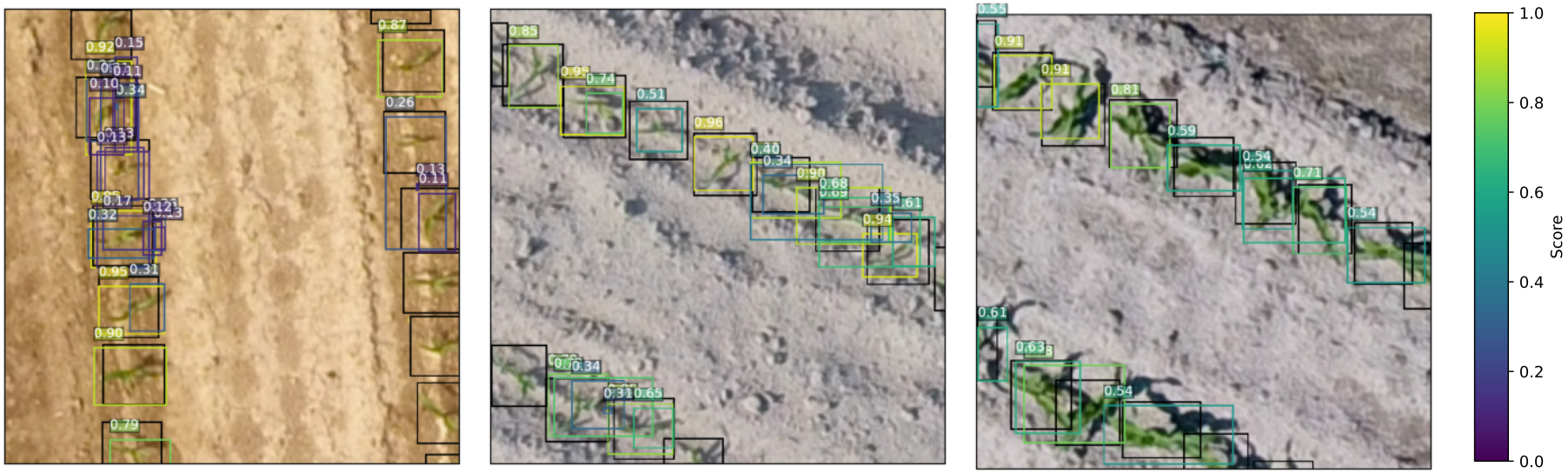


Figure 10: 50 shots CD-ViTO with ViT-B backbone predictions on the 1, 2 and 3 ID test datasets tile examples, respectively from the left hand side to the right. Black bounding boxes are the ground truth annotations, while the bounding boxes in viridis color scale are the model predictions.

also considerably elevated, ranging from around 40 to 140. Furthermore, the mAP values were lower than those of the many and few-shots 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-shots model (50-shots ViT-B backbone). Some rare case of good predictions were even more accurate than the few-shots best performance, as shown in Figure 12.

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. [22] and Andvaag et al. [13], 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.

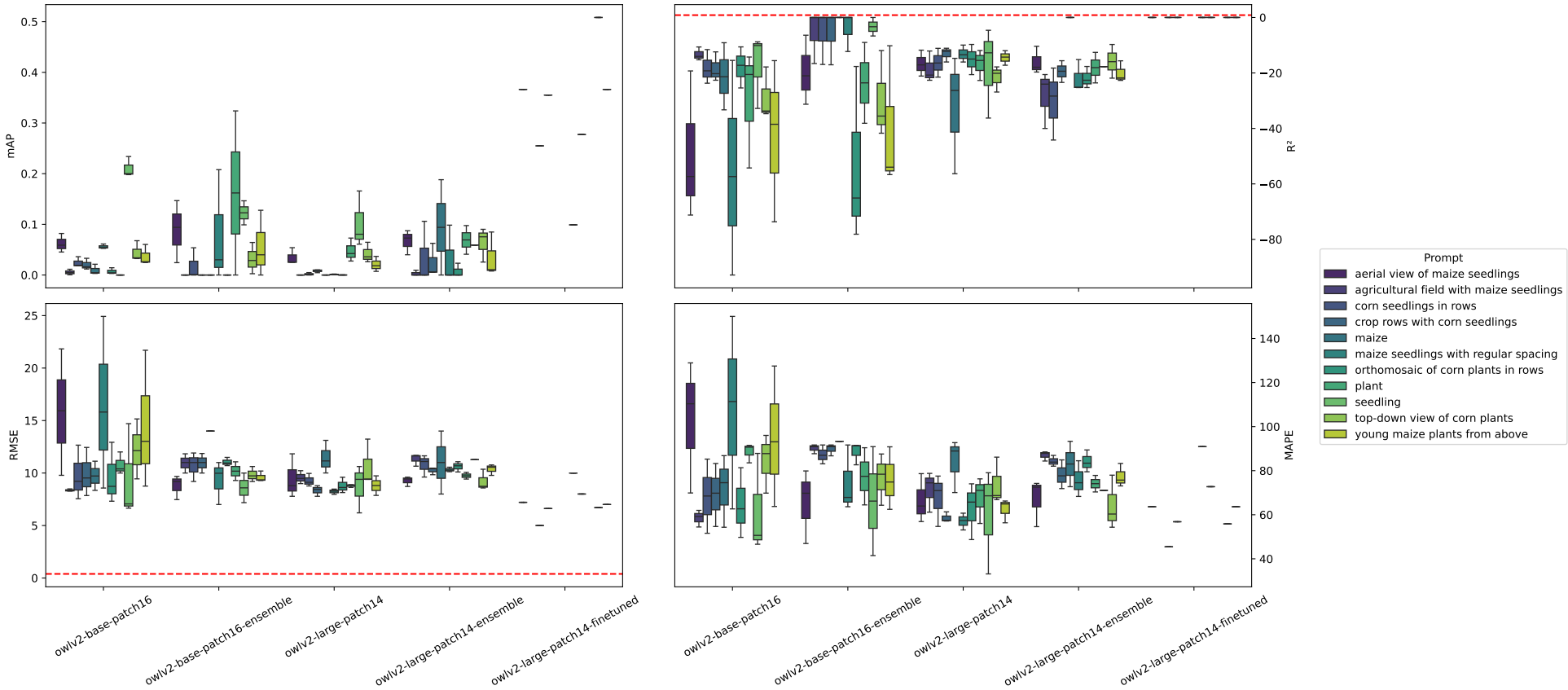


Figure 11: The figure shows the relationship between the OWLv2 model size, used prompt and model performance. The x-axis represents the model settings and the model size. Colors represent the different prompts used. The four subplots show the mAP (upper left corner), R^2 (upper right corner), $RMSE$ (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 benchmark of 0.85 and 0.39.

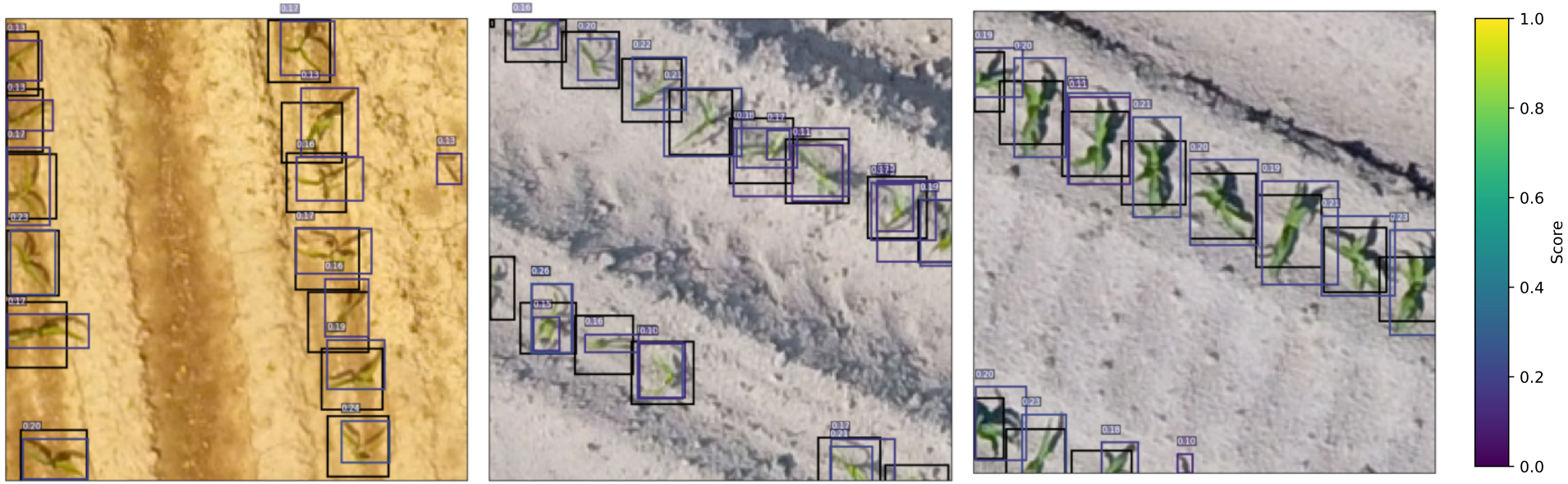


Figure 12: The best predictions with the OWLv2 model. The ID_1, ID_2 and ID_3 datasets respectively from the left hand side to the right. Prediction with owlv2-base-patch16-ensemble model of the ID_1 dataset, and with owlv2-base-patch16 model on the other two datasets. All the predictions are made with the prompt ”seedling”. Black bounding boxes are the ground truth annotations, while the bounding boxes in viridis color scale are the model predictions.

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 [15]. 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 even 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. Within YOLO-family models, we observed that the lightweight YOLOv5n, YOLOv5s, and YOLOv8n achieved the benchmark with 130, 130, and 110 sam-

ples respectively. As already well-known, increasing model complexity in CNN-based architectures corresponded to increased dataset size requirements. Conversely, for transformer-mixed models like RT-DETR, we observed different patterns, with RT-DETR L achieving the benchmark with only 60 samples while the larger RT-DETR X required 100 samples. 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. 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, that could be a limitation for some applications. 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).

The predictability of model performance based on dataset size (as evidenced by GoF values exceeding 0.3 for successful models) pro-

vides practical guidance for practitioners. 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, which can help inform efficient resource allocation for annotation efforts.

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.

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.

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 moderately-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.

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 precise maize seedling detection. The best CD-ViTO few-shot model achieved a $RMSE$ 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 [79, 62]. 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 **??**), 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 11) 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. Addressing these challenges through domain-specific adaptations could help narrow the performance gap, potentially making 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-shots 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.

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.

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 filter 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 applications, 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-shots, few-shots, and zero-shots models are occasionally able to produce annotations with sufficient quality for training ID many-shots 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 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.

Bibliography

[1] [1409.0575] ImageNet Large Scale Visual Recognition Challenge.

[2] Estimates of Maize Plant Density from UAV RGB Images Using Faster-RCNN Detection Model: Impact of the Spatial Resolution.

[3] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks | IEEE Journals & Magazine | IEEE Xplore.

[4] Maize-seedling-detection Dataset $>$ Overview.

[5] Maize_seeding Dataset $>$ Overview.

[6] Meta-Learning-Based Incremental Few-Shot Object Detection | IEEE Journals & Magazine | IEEE Xplore.

[7] You Only Look Once: Unified, Real-Time Object Detection | IEEE Conference Publication | IEEE Xplore.

[8] PP 1/333 (1) Adoption of digital technology for data generation for the efficacy evaluation of plant protection products. *EPPO Bulletin*, page epp.13037, November 2024.

[9] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 966–970, October 2022.

[10] Khaled Alhazmi, Walaa Alsumari, Indrek Seppo, Lara Podkuiko, and Martin Simon. Effects of annotation quality on model performance. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 063–067, April 2021.

[11] Hamed Amini Amirkolaee, Miaojing Shi, Lianghua He, and Mark Mulligan. AdaTreeFormer: Few shot domain adaptation for tree counting from a single high-resolution image. *ISPRS Journal of Photogrammetry and Remote Sensing*, 214:193–208, August 2024.

[12] Ayoub Benali Amjoud and Mustapha Amrouch. Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review. *IEEE access : practical innovations, open solutions*, 11:35479–35516, 2023.

[13] Erik Andvaag, Kaylie Krys, Steven J. Shirtliffe, and Ian Stavness. Counting Canola: Toward Generalizable Aerial Plant Detection Models. *Plant phenomics (Washington, D.C.)*, 6:0268, November 2024.

[14] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data Augmentation Generative Adversarial Networks, March 2018.

[15] Chetan M Badgujar, Alwin Poulose, and Hao Gan. Agricultural object detection with You Only Look Once (YOLO) Algorithm: A bibliometric and systematic literature review. *Computers and Electronics in Agriculture*, 223:109090, August 2024.

[16] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-Shot Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[17] Abel Barreto, Philipp Lottes, Facundo Ramón Ispizua Yamati, Stephen Baumgarten, Nina Anastasia Wolf, Cyrill Stachniss, Anne-Katrin Mahlein, and Stefan Paulus. Automatic UAV-based counting of seedlings in sugar-beet field and extension to maize and strawberry. *Computers and Electronics in Agriculture*, 191:106493, December 2021.

[18] L. Brigato and L. Iocchi. A Close Look at Deep Learning with Small Data, October 2020.

[19] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers, May 2020.

[20] Clément Chadebec and Stéphanie Allassonnière. Data Augmentation with Variational Autoencoders and Manifold Sampling. In Sandy Engelhardt, Ilkay Oksuz, Dajiang Zhu, Yixuan Yuan, Anirban Mukhopadhyay, Nicholas Heller, Sharon Xiaolei Huang, Hien Nguyen, Raphael Sznitman, and Yuan Xue, editors, *Deep Generative Models, and Data Augmentation,*

Labelling, and Imperfections, pages 184–192, Cham, 2021. Springer International Publishing.

[21] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning Augmentation Policies from Data, April 2019.

[22] Etienne David, Gaëtan Daubige, François Joudelat, Philippe Burger, Alexis Comar, Benoit de Solan, and Frédéric Baret. 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 sunflowe, 2021.

[23] 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.

[24] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V. Le, and Xiaodan Song. SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2020.

[25] FAO. In *Agricultural Production Statistics 2010–2023*, volume Analytical Briefs. FAOSTAT, Rome, 2024.

[26] Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In Martin A. Fischler and Oscar Firschein, editors, *Readings in Computer Vision*, pages 726–740. Morgan Kaufmann, San Francisco (CA), January 1987.

[27] Kun Fu, Tengfei Zhang, Yue Zhang, Menglong Yan, Zhonghan Chang, Zhengyuan Zhang, and Xian Sun. Meta-SSD: Towards Fast Adaptation for Few-Shot Object Detection With Meta-Learning. *IEEE access : practical innovations, open solutions*, 7:77597–77606, 2019.

[28] Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-Domain Few-Shot Object Detection via Enhanced Open-Set Object Detector, September 2024.

[29] Héctor García-Martínez, Héctor Flores-Magdaleno, Abdul Khalil-Gardezi, Roberto Ascencio-Hernández, Leonardo Tijerina-Chávez, Mario A. Vázquez-Peña, and Oscar R. Mancilla-Villa. Digital Count of Corn Plants Using Images Taken by Unmanned Aerial Vehicles and Cross Correlation of Templates. *Agronomy*, 10(4):469, April 2020.

[30] Tingting Geng, Haiyang Yu, Xinru Yuan, Ruopu Ma, and Pengao Li. Research on Segmentation Method of Maize Seedling Plant Instances Based on UAV Multispectral Remote Sensing Images. *Plants*, 13(13):1842, January 2024.

[31] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the Backbones: A Large-Scale Comparison of Pretrained Models across Computer Vision Tasks, November 2023.

[32] Nico Heider, Lorenz Gunreben, Sebastian Zürner, and Martin Schieck. A Survey of Datasets for Computer Vision in Agriculture: A catalogue of high-quality RGB image datasets of natural field scenes. *45. GIL-Jahrestagung, Digitale Infrastrukturen für eine nachhaltige Land-, Forst und Ernährungswirtschaft*, pages 35–47, 2025.

[33] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization, June 2020.

[34] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically, December 2017.

[35] Gabriel Huang, Issam Laradji, David Vazquez, Simon Lacoste-Julien, and Pau Rodriguez. A Survey of Self-Supervised and Few-Shot Object Detection, August 2022.

[36] Pranav Jeevan and Amit Sethi. Which Backbone to Use: A

Resource-efficient Domain Specific Comparison for Computer Vision, June 2024.

[37] Yu Jiang, Changying Li, Andrew H. Paterson, and Jon S. Robertson. DeepSeedling: Deep convolutional network and Kalman filter for plant seedling detection and counting in the field. *Plant Methods*, 15(1):141, November 2019.

[38] Longlong Jing and Yingli Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey, February 2019.

[39] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. GitHub Ultralytics YOLO, January 2023.

[40] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-Shot Object Detection via Feature Reweighting, October 2019.

[41] Azam Karami, Melba Crawford, and Edward J. Delp. Automatic Plant Counting and Location Based on a Few-Shot Learning Technique. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5872–5886, 2020.

[42] Sushma Katari, Sandeep Venkatesh, Christopher Stewart, and Sami Khanal. Integrating Automated Labeling Framework for Enhancing Deep Learning Models to Count Corn Plants Using UAS Imagery. *Sensors*, 24(19):6467, January 2024.

[43] Asifullah Khan, Zunaira Rauf, Anabia Sohail, Abdul Rehman Khan, Hifsa Asif, Aqsa Asif, and Umair Farooq. A survey of

the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, 56(3):2917–2970, December 2023.

[44] Rahima Khanam and Muhammad Hussain. YOLOv11: An Overview of the Key Architectural Enhancements, October 2024.

[45] Bruno T. Kitano, Caio C. T. Mendes, André R. Geus, Henrique C. Oliveira, and Jefferson R. Souza. Corn Plant Counting Using Deep Learning and UAV Images. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2019.

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

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[48] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[49] Wenwen Li and Yun Zhang. DC-YOLO: An improved field plant detection algorithm based on YOLOv7-tiny. *Scientific Reports*, 14(1):26430, November 2024.

[50] Yang Li, Zhiyuan Bao, and Jiangtao Qi. Seedling maize counting method in complex backgrounds based on YOLOV5 and

Kalman filter tracking algorithm. *Frontiers in Plant Science*, 13, November 2022.

[51] Yong Li, Naipeng Miao, Liangdi Ma, Feng Shuang, and Xingwen Huang. Transformer for object detection: Review and benchmark. *Engineering Applications of Artificial Intelligence*, 126:107021, November 2023.

[52] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to Learn Quickly for Few-Shot Learning, September 2017.

[53] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection, February 2018.

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015.

[55] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 38–55, Cham, 2025. Springer Nature Switzerland.

[56] Shuaibing Liu, Dameng Yin, Haikuan Feng, Zhenhai Li, Xiaobin Xu, Lei Shi, and Xiuliang Jin. Estimating maize seedling

number with UAV RGB images and advanced image processing methods. *Precision Agriculture*, 23(5):1604–1632, October 2022.

[57] Wenxin Liu, Jing Zhou, Biwen Wang, Martin Costa, Shawn M. Kaeppler, and Zhou Zhang. IntegrateNet: A Deep Learning Network for Maize Stand Counting From UAV Imagery by Integrating Density and Local Count Maps. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[58] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. TasselNet: Counting maize tassels in the wild via local counts regression network. *Plant Methods*, 13(1):79, November 2017.

[59] Mélissande Machefer, François Lemarchand, Virginie Bonnefond, Alasdair Hitchins, and Panagiotis Sidiropoulos. Mask R-CNN Refitting Strategy for Plant Counting and Sizing in UAV Imagery. *Remote Sensing*, 12(18):3015, January 2020.

[60] Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sania Fidler, and Marc T. Law. How Much More Data Do I Need? Estimating Requirements for Downstream Tasks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 275–284, New Orleans, LA, USA, June 2022. IEEE.

[61] Uwe Meier, Hermann Bleiholder, Liselotte Buhr, Carmen Feller, Helmut Hack, Martin Heß, Peter D. Lancashire, Uta Schnock,

Reinhold Stauß, Theo van den Boom, Elfriede Weber, and Peter Zwerger. The BBCH system to coding the phenological growth stages of plants – history and publications –. *Journal für Kulturpflanzen*, 61(2):41–52, February 2009.

[62] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, December 2023.

[63] Samuel G. Müller and Frank Hutter. TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation, August 2021.

[64] Nhat-Duy Nguyen, Tien Do, Thanh Duc Ngo, Duy-Dinh Le, and Cesare F. Valenti. An Evaluation of Deep Learning Methods for Small Object Detection. *JECE*, 2020, January 2020.

[65] 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.

[66] Aref Miri Rekavandi, Shima Rashidi, Farid Boussaid, Stephen Hoefs, Emre Akbas, and Mohammed bennamoun. Transform-

ers in Small Object Detection: A Benchmark and Survey of State-of-the-Art, September 2023.

[67] Gianmarco Roggiolani, Federico Magistri, Tiziano Guadagnino, Jan Weyler, Giorgio Grisetti, Cyrill Stachniss, and Jens Behley. On Domain-Specific Pre- Training for Effective Semantic Perception in Agricultural Robotics. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11786–11793, May 2023.

[68] Liangbing Sa, Chongchong Yu, Zhaorui Hong, Tong Zheng, and Sihan Liu. A broader study of cross-domain few-shot object detection. *Applied Intelligence*, 53(23):29465–29485, December 2023.

[69] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019.

[70] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, August 2017.

[71] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection, July 2020.

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Process-*

ing Systems, NIPS'17, pages 6000–6010, Red Hook, NY, USA, December 2017. Curran Associates Inc.

[73] K. Velumani, R. Lopez-Lozano, S. Madec, W. Guo, J. Gillet, A. Comar, and F. Baret. Estimates of Maize Plant Density from UAV RGB Images Using Faster-RCNN Detection Model: Impact of the Spatial Resolution. *Plant phenomics (Washington, D.C.)*, 2021:9824843, January 2021.

[74] Biwen Wang, Jing Zhou, Martin Costa, Shawn M. Kaeppler, and Zhou Zhang. Plot-Level Maize Early Stage Stand Counting and Spacing Detection Using Advanced Deep Learning Algorithms Based on UAV Imagery. *Agronomy*, 13(7):1728, July 2023.

[75] Dongxue Wang, Rajamohan Parthasarathy, and Xian Pan. 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*, ICSCIS '24, pages 635–639, New York, NY, USA, August 2024. Association for Computing Machinery.

[76] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[77] Kesheng Wu, Ekow Otoo, and Arie Shoshani. Optimizing connected component labeling algorithms. January 2005.

[78] Xiongwei Wu, Doyen Sahoo, and Steven Hoi. Meta-RCNN: Meta Learning for Few-Shot Object Detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 1679–1687, New York, NY, USA, October 2020. Association for Computing Machinery.

[79] Guanyu Xu, Zhiwei Hao, Yong Luo, Han Hu, Jianping An, and Shiwen Mao. DeViT: Decomposing Vision Transformers for Collaborative Inference in Edge Devices, September 2023.

[80] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. *Meta-DETR: Few-Shot Object Detection via Unified Image-Level Meta-Learning*. March 2021.

[81] Song Zhang, Yehua Yang, Lei Tu, Tianling Fu, Shenxi Chen, Fulang Cen, Sanwei Yang, Quanzhi Zhao, Zhenran Gao, and Tengbing He. Comparison of YOLO-based sorghum spike identification detection models and monitoring at the flowering stage. *Plant Methods*, 21(1):20, February 2025.

[82] Kai Zhao, Lulu Zhao, Yanan Zhao, and Hanbing Deng. Study on Lightweight Model of Maize Seedling Object Detection Based on YOLOv7. *Applied Sciences*, 13(13):7731, January 2023.

[83] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs Beat YOLOs on Real-time Object Detection, April 2024.

[84] Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with Collaborative Hybrid Assignments Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6725–6735, Paris, France, October 2023. IEEE.

[85] Hongwei Zou, Hao Lu, Yanan Li, Liang Liu, and Zhiguo Cao. Maize tassels detection: A benchmark of the state of the art. *Plant Methods*, 16(1):108, August 2020.

2.2 Ordinal and Nominal Variables

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]	Image processing	Not reported	Involving only biomass estimation,
no AI involved, and no monitorable stability			
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.

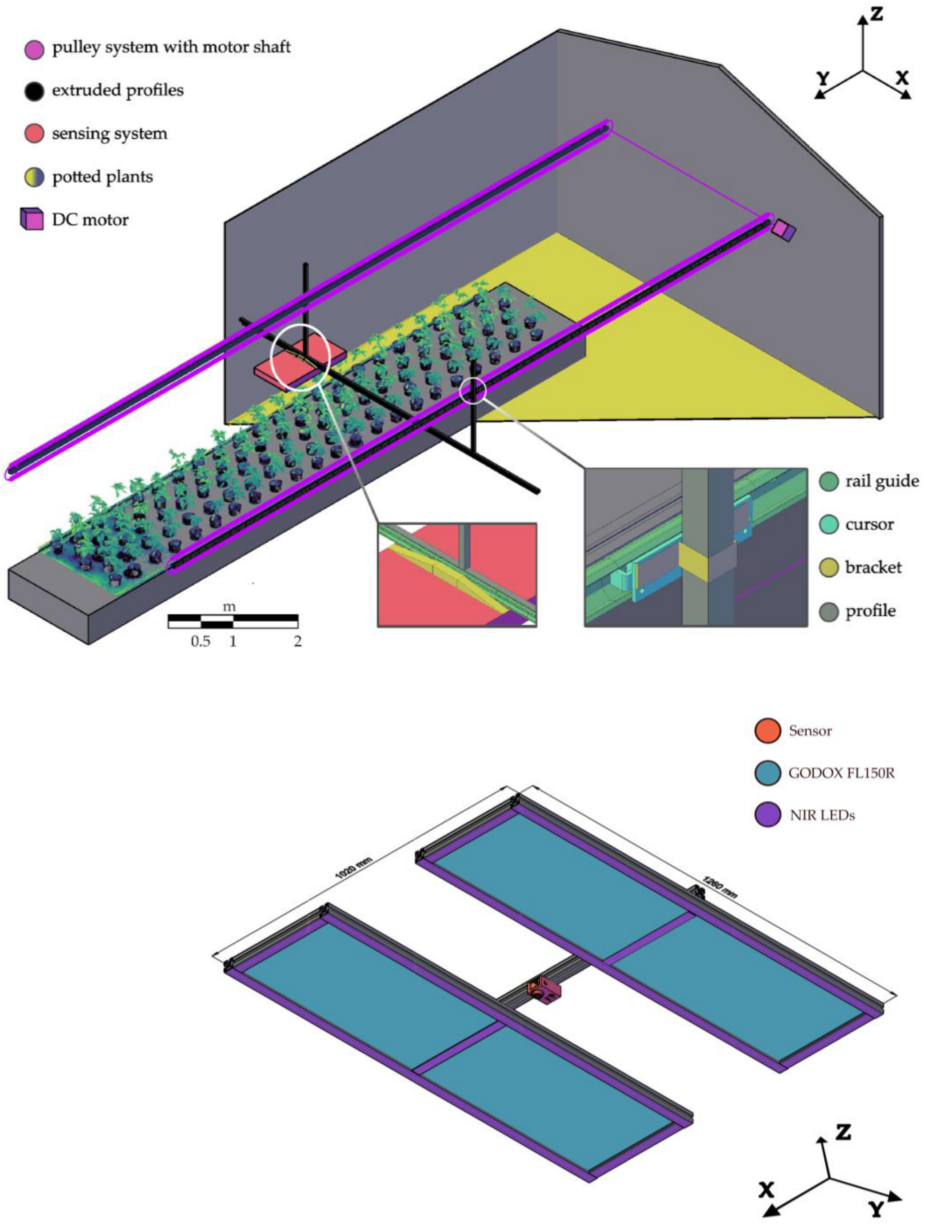


Figure 1: Platform and sensing system (top) and only the sensing system (bottom).

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.

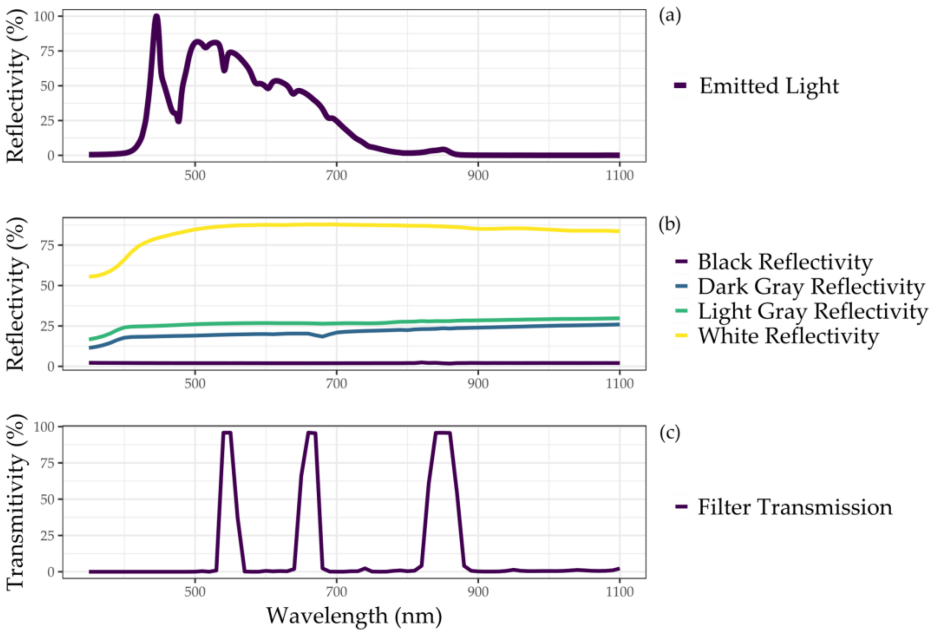


Figure 2: (a) Spectral signature of the reference panel, lighted with the tested LEDs and measured using the RS-5400 Spectro-radiometer. (b) Reflectivity of the MAPIR calibration panels corresponding to the different grayscale levels (yellow, light-green, blue, and violet colors in the graph) provided 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.

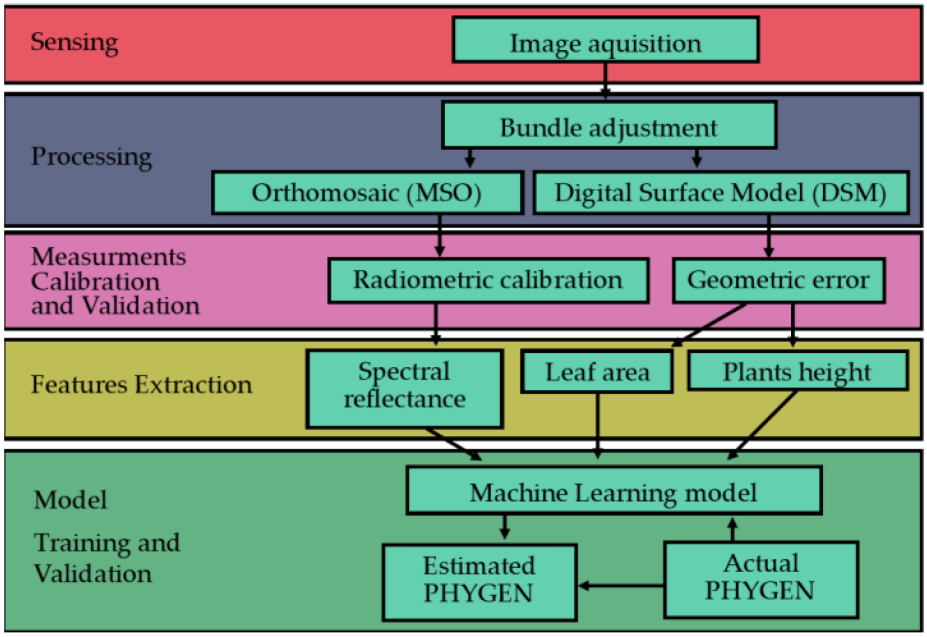


Figure 3: General workflow of the suggested method.

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.

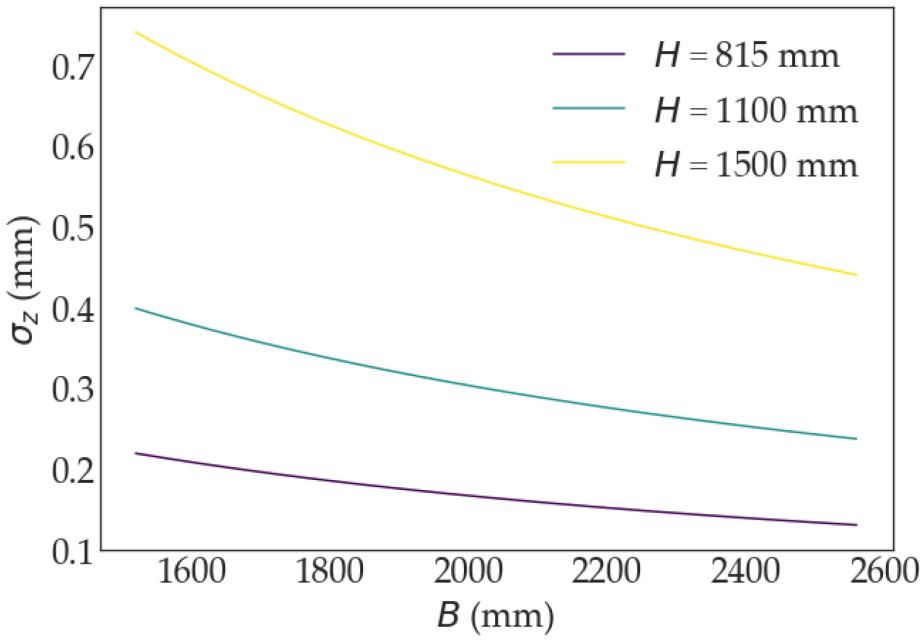


Figure 4: σ_z estimates computed by eq. (2). Colored curves refer to different D values.

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.

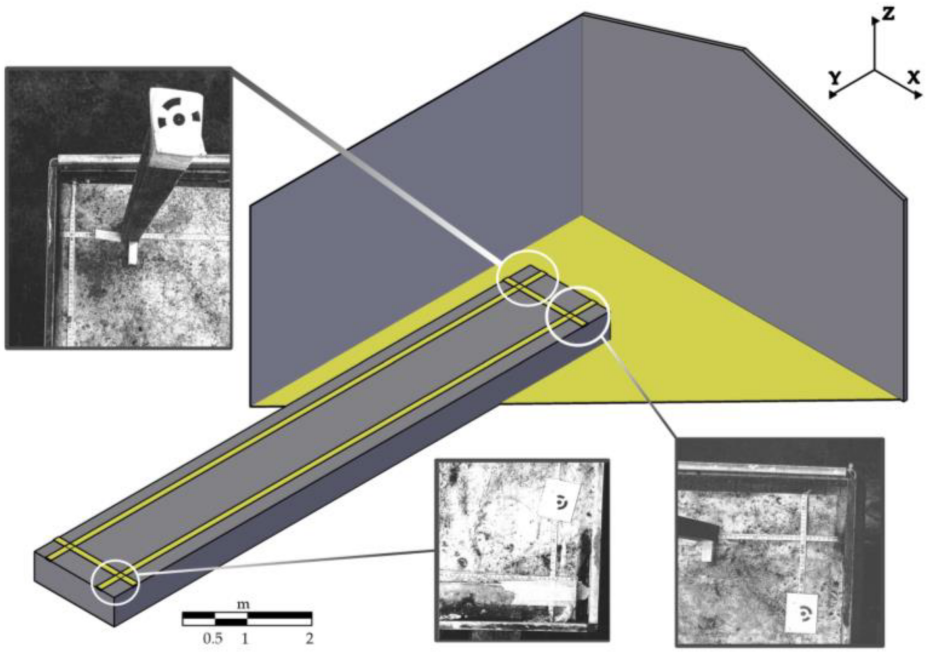


Figure 5: Metered tapes and GCPs on the bench.

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.

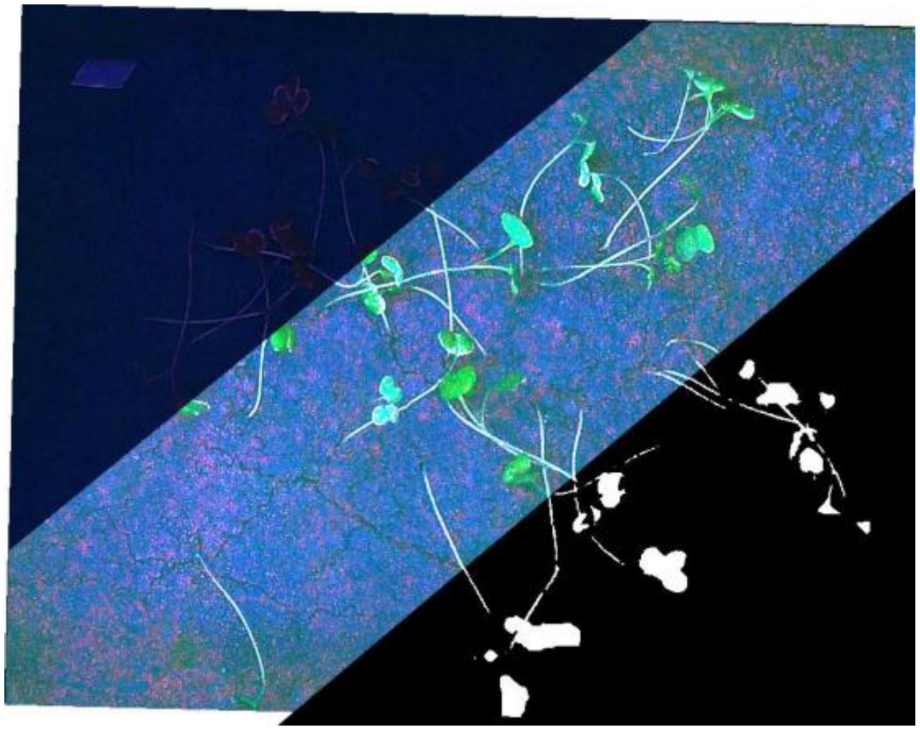


Figure 6: Non-calibrated (dark) and calibrated MSO are shown together with the last vegetation mask (white pixels) on an oil seed rape pot.

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)	$\sum_{i=1}^{n}\left(y_i-\frac{L}{1+e^{-k(\hat{y}_i-y_0)}}\right)$ 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.

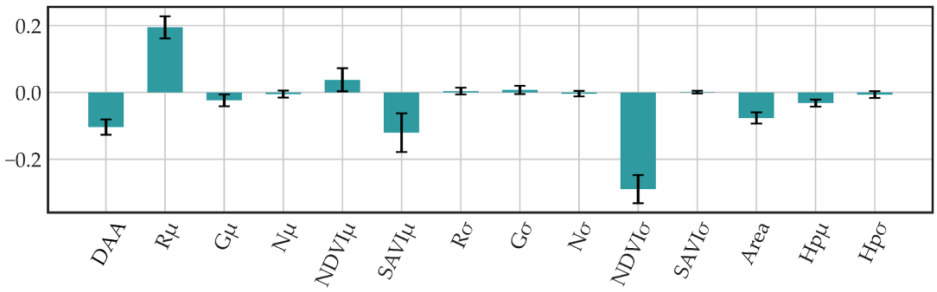


Figure 7: Mean values of LASSO β coefficients from the 10-fold approach, given for all the predictors. Whisker bars show 1-sigma LASSO β estimates.

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.

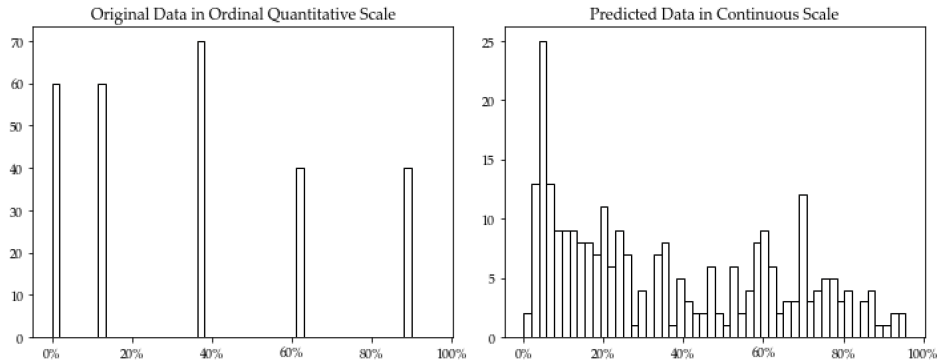


Figure 8: Discrete PHYGEN scores from the ordinary human vision-based approach (left). Continuous PHYGEN scores from the model proposed in this work (right).

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. https://drive.google.com/file/d/10gIzOjWVNoG9dvZwmAUG9fVqkEZHXEur/view?usp=drive_op

[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_meetings/https%3A%2F%2Fwww.eppo.int%2FMEETIN

[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.

2.3 Binary Variables

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 [17]. Early and accurate detection of plant diseases is crucial for implementing timely interventions, reducing pesticide use, and preventing disease spread across agricultural landscapes [13]. 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 [7, 14]. Deep learning approaches, particularly convolutional neural networks (CNNs) and vision transformers, have demonstrated remarkable success in classifying plant diseases from leaf images [19]. 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 [23].

Anomaly detection presents a promising alternative paradigm that requires training only on healthy samples, identifying diseased specimens as deviations from the normal state [16, 5]. This approach aligns well with agricultural monitoring scenarios where healthy plants constitute the majority class, and various diseases represent anoma-

lous conditions [11]. Additionally, anomaly detection frameworks can potentially identify novel or previously unseen disease manifestations that supervised classifiers would struggle to recognize [4].

A critical challenge in developing robust plant disease detection systems is the significant performance gap between controlled laboratory environments and real-world field conditions [22]. 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 [10] 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 [21], 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.

2.3.2.2 Tested Backbones

This study evaluates a comprehensive set of neural network architectures as feature extractors for anomaly detection. We imple-

mented 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 [8]. 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 [18].
- **DenseNet family:** DenseNet121, DenseNet161, DenseNet169, DenseNet201, featuring dense connectivity patterns that strengthen feature propagation [9].
- **Efficient architectures:** EfficientNet (B0-B7), EfficientNetV2 (S, M, L), MobileNetV2, MobileNetV3, which are optimized for computational efficiency while maintaining high accuracy [20].
- **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 [6].
- **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 [12].

- **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 [15].

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 coordinate 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 [**?**] 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 [**?**] 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 [**?**] 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 1 and Figure 2 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 3 and Figure 4 show the distribution of clustering accuracy across all tested configurations respectively for the Plant Village and the Plant Pathology datasets.

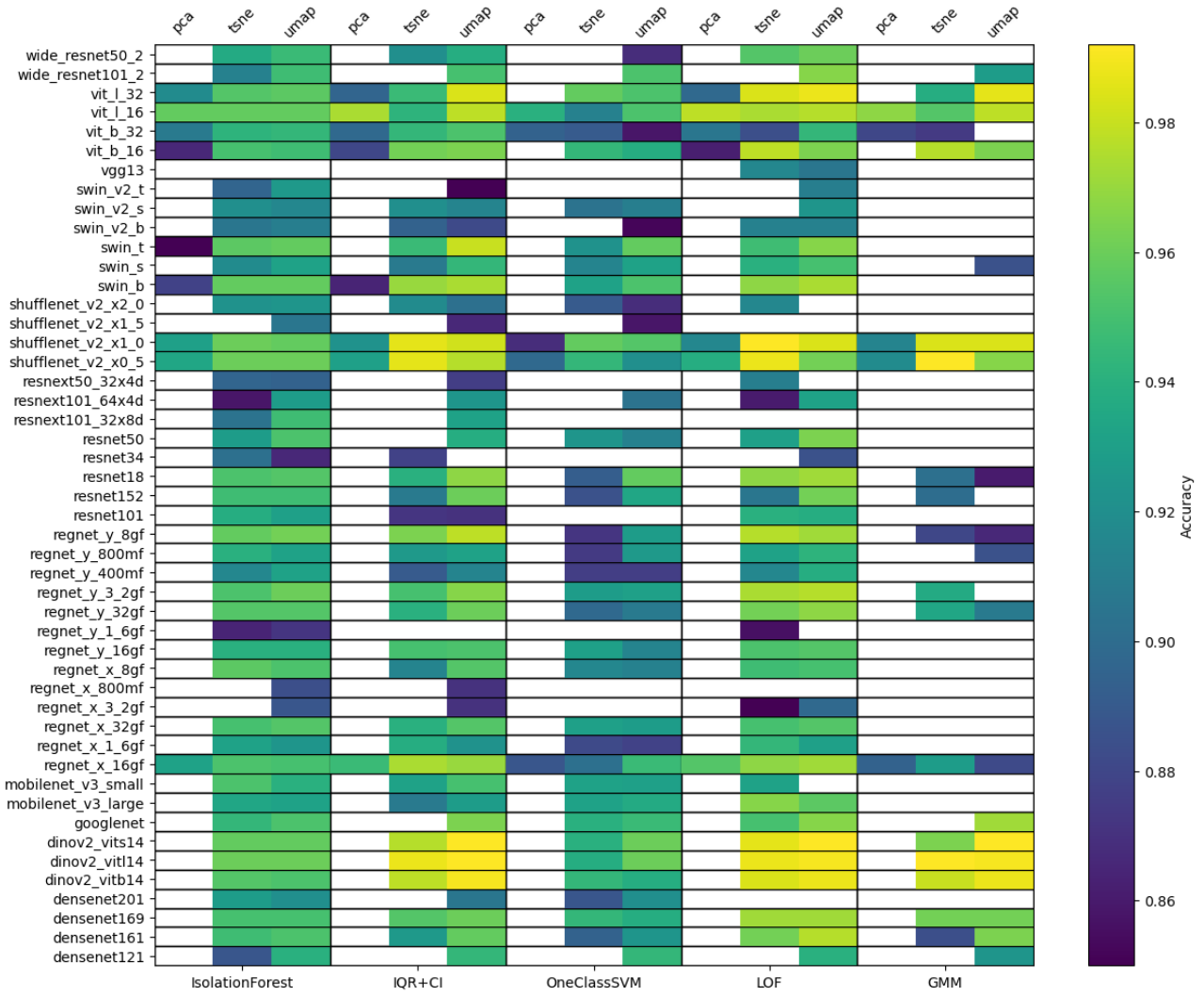


Figure 1: Anomaly detection performance across different backbone architectures and dimensionality reduction techniques on the Plant Village dataset. Backbones on y-axis, anomaly detetection algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for each backbone-detection-reduction combination.

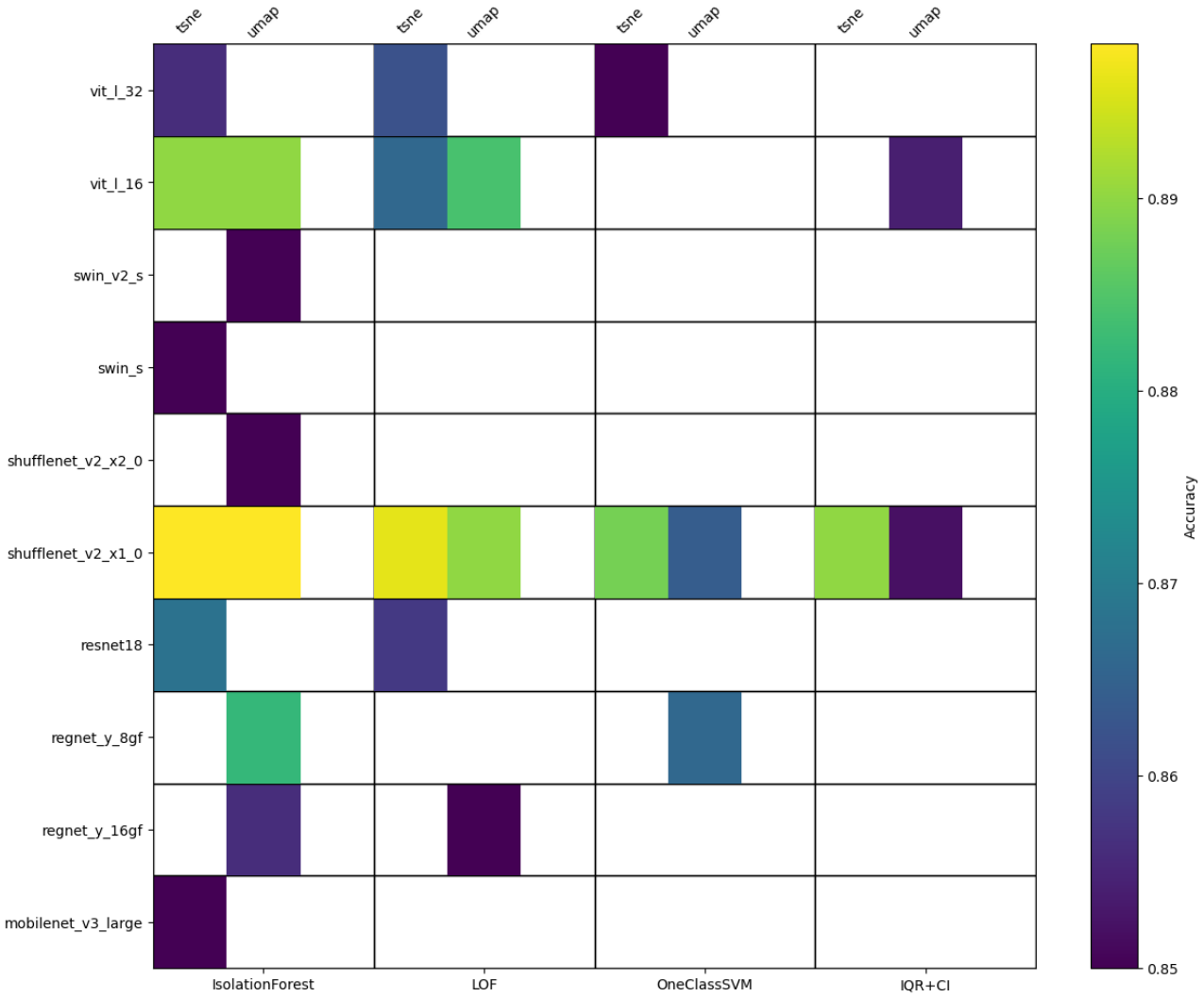


Figure 2: Anomaly detection performance across different backbone architectures and dimensionality reduction techniques on the Plant Pathology dataset. Backbones on y-axis, anomaly detetection algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for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

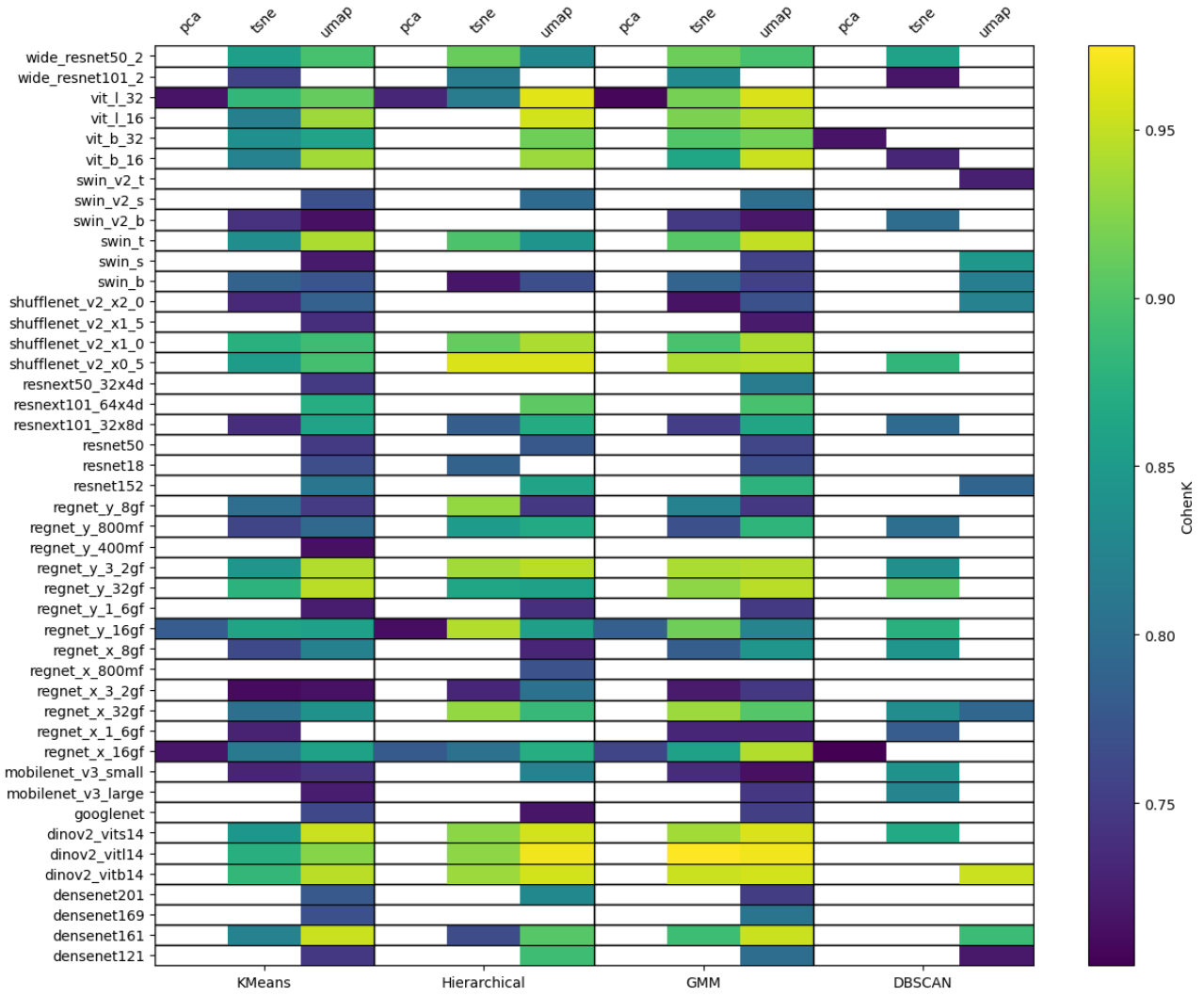


Figure 3: Clustering performance across different backbone architectures and dimensionality reduction techniques on the Plant Village dataset. Backbones on y-axis, clustering algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for each backbone-detection-reduction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

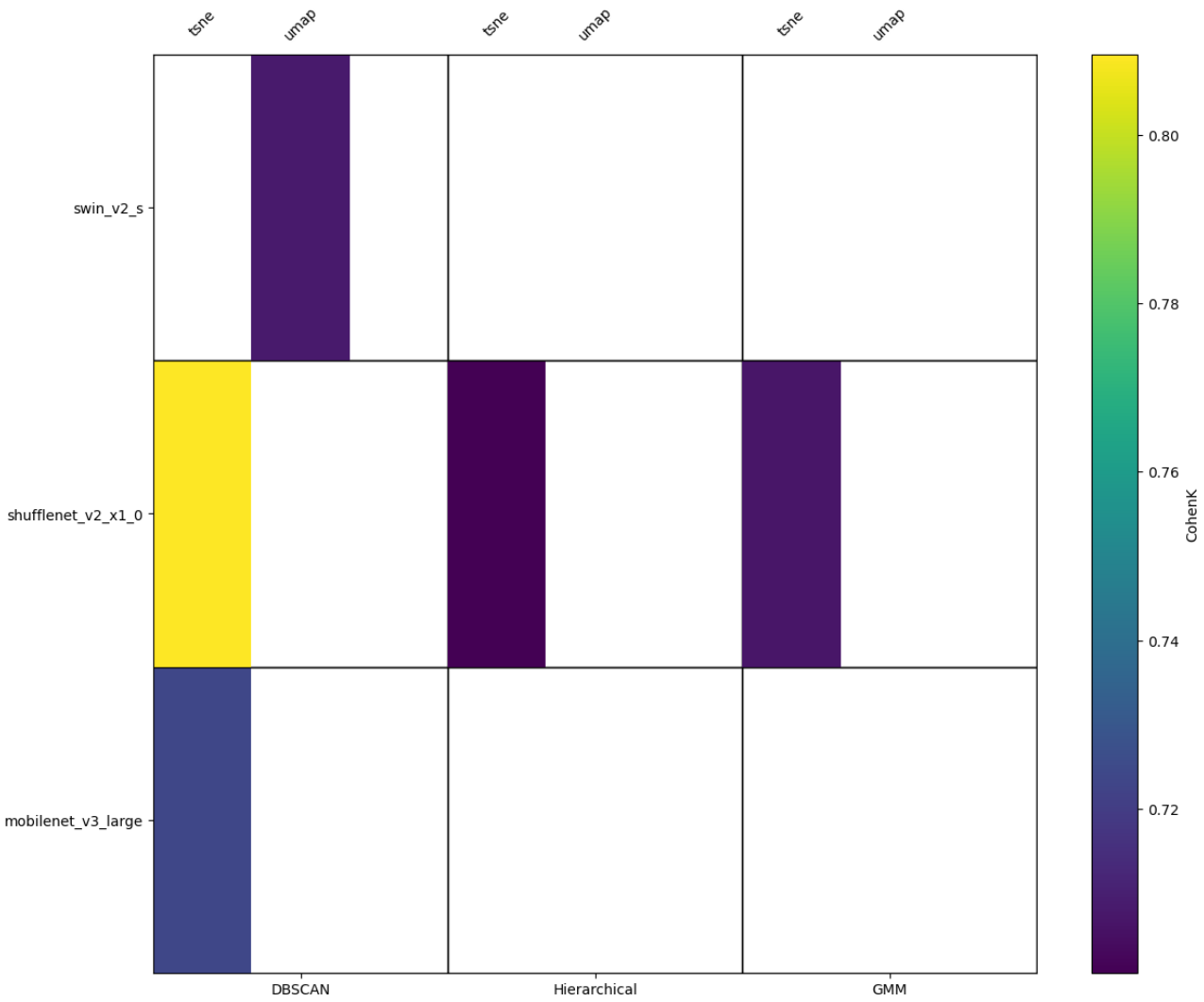


Figure 4: Clustering performance across different backbone architectures and dimensionality reduction techniques on the Plant Pathology dataset. Backbones on y-axis, clustering algorithm on lower x-axis, dimensionality reduction method on top x-axis. The color indicates the accuracy for each backbone-detection-reduction 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 1. 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 2.

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 3 and 4. 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] 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.

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

[6] 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].

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

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

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

[10] 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].

[11] 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].

[12] 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].

[13] 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.

[14] 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.

[15] 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].

[16] 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].

[17] 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.

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

[19] 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.

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

[21] 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].

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

[23] 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.