sciendo

CAIM
ISSN 2038-0909

Commun. Appl. Ind. Math. 15 (2), 2024, 69–90 DOI: 10.2478/caim-2024-0014

Research Article

Exploratory Data Analysis and Supervised Learning in Plant Phenotyping Studies

Vincenzo Schiano Di Cola[1*], Mariachiara Cangemi[2,3*], Simone Scala[4*], Stephan Summerer[5], Maurilia Maria Monti[3], Francesco Loreto[2,3], Salvatore Cuomo[4]

[1]Water research institute (IRSA), National Research Council (CNR), Bari, Italy

[2]Department of Biology, University of Naples Federico II, Napoli, Italy

[3]Institute for Sustainable Plant Protection (IPSP) National Research Council (CNR), Napoli, Italy

[4]Department of Mathematics and Applications "R. Caccioppoli", University of Naples Federico II, Napoli, Italy

[5]ALSIA Centro Ricerche Metapontum Agrobios, Metaponto, Italy

[*]Email address for correspondence: vincenzo.schianodicola@ba.irsa.cnr.it, mariachiara.cangemi@unina.it

Communicated by Clemente Cesarano

Received on 07 22, 2024. Accepted on 10 21, 2024.

Abstract

This study investigates the use of exploratory data analysis and supervised learning techniques to analyze plant phenotyping traits, with a specific focus on: i) genetic diversity (wild type vs mutant tomato plants); ii) plant-plant interactions (primed vs non-primed plants using volatiles emitted by other stressed plants); and iii) plant stress response (using drought stress and comparing droughted plants with controls). The analyzed data consisted of high-throughput imaging at multiple wavelengths, which allowed for the examination of various morphological traits. The dataset contained the phenotypic characteristics of both wildtype and mutated tomato plants exposed to water stress. Machine learning algorithms were used to identify significant phenotypic indicators and predict plant stress responses. The use of techniques such as K-means clustering and Bayesian classifiers provided valuable insights into the temporal dynamics of plant traits under a variety of experimental conditions. This research emphasizes the importance of employing advanced statistical and machine learning methods to improve the precision and efficacy of phenotypic analysis in plant sciences.

Keywords: `K-means clustering; Bayesian classifiers; Statistical data analysis; Plant stress response; Machine learning in plant science`

AMS subject classification: `62H30; 92C80; 68T05`

1. Introduction

Plant phenotyping is a discipline of growing interest in agriculture and plant sciences. It is based on the study, measurement, and non-destructive analysis of morpho-physiological characteristics of plants to understand the relationships between the genotype, i.e. the genetic (hereditary) background of an individual, and the environment. In fact, the phenotype is the result of the genotype x environment interaction, with farming practices also to be considered as a third factor in the case of cultivated plants [1].

Image analysis is a high-throughput tool of great importance in plant phenotyping, with growing applications. Imaging in the visible part of the light spectrum (400-700 nm) is used to measure plant architecture parameters, such as total and projected biomass, leaf area and color, root growth and architecture, fruit number and growth. Fluorescence imaging has been predominantly used for monitoring photosynthetic efficiency and the onset and development of abiotic and biotic stresses that impair photosynthesis. Infrared thermal imaging can be used to assess plant and soil temperature, also inferring the onset and development of water stress, and the consequent stomatal closure in plants [2]. When used throughout plant growth and development, image analysis is a powerful tool to accelerate breeding programs, improve environmental resource management, and contribute to the generation of predictive

models of plant response to climate change [3]. Collected images are processed and analyzed using computer vision and machine learning software, allowing automation of recognition and quantification of phenotypic traits, and improved accuracy and speed of the analyses.

1.1. *An application of image analysis for studying plant-plant communication.*

In this study, we used images of plant phenotypes to investigate plant-plant communication via the release of volatile organic compounds (VOCs). VOCs released by plants are known to allow communication between plants and other organisms, and the VOCs sensing mechanisms in insects and mammals are well-known. VOCs may also allow plant communication with other plants, but how plants sense VOCs is unclear [4].

VOCs can affect genetic regulation, metabolism, physical characteristics, response to stress, and behavioral choices in recipient plants, via a mechanism known as “priming”. Primed plants show more efficient responses to additional stimuli, as well as increased resistance or tolerance or resilience to stresses [5]. We aimed at detecting possible plant phenotypic markers associated to VOC-sensing in plants.

An experiment was conducted on two lines of *Solanum lycopersicum* (tomato) plants: a wildtype line (WT) and a mutated line (MT). The MT line is affected by a mutation in a region coding for odorant receptors (OR) associated with odorant-binding proteins (OBP). The experimental design consisted of the following phases: WT and MT line plants were placed in closed containers with tomato plants previously subjected to water stress, acting as ”VOC emitters”. After 24 hours of exposure (”priming”), some plants from both lines were subjected to water stress. Data as images were collected and machine learning algorithms were used to study and compare responses to water stress, from non-primed, stressed WT, and MT plants, searching for insights and predictors to design new experiments and start to build a model.

1.2. *Mathematical contribution*

The increasing amount of agricultural data poses both advantages and difficulties. Machine learning (ML) approaches provide effective tools for harnessing the potential of this data, allowing to automate processes, enhance predictions, optimize agricultural practices to meet increasing demands while guaranteeing sustainability, extract significant patterns, and make accurate choices. In recent years, the integration of machine learning, optimization methods, and inverse problem solutions has become a prominent research focus in agricultural data analysis. Models that efficiently learn from data and make accurate predictions in biology, finance, and engineering require this integration. Machine learning is dependent on optimization for the purposes of algorithm training and model tuning. Gradient descent, Newton’s technique, and its derivatives are employed to minimize loss functions and enhance model precision. The process of optimization has enhanced the analysis of gene expression data by converting continuous data into discrete values and enhancing the performance of machine learning models [6]. In medical imaging and geophysics, inverse problems entail deducing system parameters from observed data. The solving of machine learning inverse problems requires the application of complex algorithms adept at handling noisy and incomplete data. By integrating inverse problem methodologies with machine learning, models can uncover hidden structures within high-dimensional data. This is particularly beneficial in bioinformatics, where gene expression data often includes noise and possesses a high dimensionality [7].

This study explores several major machine learning approaches, beginning with pre-processing tools, K-means clustering and Bayesian classifiers. These techniques are used to extract significant insights from data, such as markers or feasible predictors finding the subsets of the most important features in high-dimensional spaces. Data analysis requires pre-processing, this involves handling missing values, normalizing data, and formatting it for analysis. Our dataset was obtained from image analysis as tabular, so we used Min-Max normalization, this ensures that all features contribute equally to the analysis, preventing larger range features from dominating results. Principal Component Analysis (PCA) was used to visualize and understand feature variance. PCA finds principal components, orthogonal directions that capture the most variance, to reduce data dimensionality. This step simplifies the dataset and

highlights key features, making analyses easier. Considering the limitations of time-batching, relying solely on it to identify essential features would have been inadequate. Therefore, K-means clustering was employed to explore more effective methods of distinguishing between different plant conditions. By utilizing unsupervised learning, we were able to mitigate any potential biases introduced by human interpretation of the data. In this scenario, if we assume that the data does not have pre-established categories, the act of grouping comparable data points together could have revealed hidden structures and patterns within the datasets. However, K-means clustering does not provide enough information to determine the importance of features. Bayesian classifiers, which are part of the supervised learning strategy, allow for the evaluation of feature importance by generating permutations of the features as the final result.

In order to assess the significance of each feature, we employed a non-parametric method called feature permutation importance. This approach involves changing the values of each feature and quantifying the subsequent reduction in model accuracy. Features that exhibit a substantial decrease in accuracy when rearranged are regarded as noteworthy. This methodology facilitates the identification of the most influential phenotypic characteristics that distinguish between different situations, such as plants under stress compared to those that are not stressed, or genotypes that are wildtype versus mutant. As in [8], permutations are used to correct bias in feature importance measures within Random Forest models. The methodology involves repeatedly permuting the outcome variable to generate a distribution of measured importance scores based on the null hypothesis that a feature is not informative. This method allows for the normalization of feature importance measures by comparing the observed importance to the null distribution. As a result, the corrected feature importance rankings become more accurate and interpretable, thereby improving the model's overall understanding.

The phenotypic data obtained throughout time required a time-series analysis to accurately capture the dynamic changes in plant properties. We represented each plant as a time series for each attribute, allowing us to analyze the temporal patterns and relationships among various phenotypic features. This investigation yielded valuable insights into the proliferation and maturation of plants under different experimental conditions, emphasizing crucial time intervals where noteworthy alterations were placed.

2. Materials and Methods

This part provides a full explanation of the processes involved in data collecting and analysis. It aims to ensure an adequate understanding of the experimental setup and the analytical methods employed.

2.1. *Data acquisition*

The High Throughput Plant Phenomics Platform (HTP) utilized in this study (PhenoLab) is located at the ALSIA Metapontum Agrobios Research Centre (Metaponto, Italy) and is based on a LemnaTec Scanalyzer 3D system. This system includes an automated belt conveyor and allows for precise plant identification via barcode and RFID tracking technology. The imaging setup consists of sequential camera stations capturing 3D images in near-infrared (NIR), ultraviolet (UV), and visible light (RGB). The platform also includes an automated irrigation system with a weighing station. The entire system is supported by a comprehensive digital infrastructure that manages data acquisition, storage, and processing [9].

Image acquisitions began on one-month-old plants and continued for two months, with an average of two acquisitions each week. For each plant, three images were captured: one from the top (Top View, T) and two from the sides (Side View, S) with a relative angle of 90 degrees. These images were acquired using an RGB camera with a SONY ICX274 CCD sensor [10]. The Lemna Grid software (https://www.lemnatec.com/lemnagrid-now-with-machine-learning/) that comes with the LemnaTec Scanalyzer 3D system.

2.2. *Dataset description and features*

The used dataset was in CSV file format and included comprehensive information on multiple phenotypic traits of plants. The data were gathered to investigate the impact of two treatments (treatment1:

priming; treatment2: stress) on two genotypes (wild-type and mutant). The dataset included 681 observations and 85 columns, with each entry representing a unique plant identified by a barcode (plantbarcode); 40 plants in total were measured on each day, comprising 5 replicates for every experimental thesis. The variables were categorized into temporal, genotypic, treatment and phenotypic traits.

Table 1. Description of all the variables in the dataset. Phenotypic traits are a vector of numerical variables extracted by the images.

Variable	**Description**	**Type**
Plantbarcode	Unique identifier of the plant	Text
Timestamp	Date and time of the measurement	Temporal
Date	Date of the measurement	Temporal
Genotype	Wildtype/Mutant	Categorical
Treatment1	Primed/No primed	Categorical
Treatment2	Stress/No stress	Categorical
Replica	Experiment replication	Numerical
PotId	Identifier of the pot	Text
Phenotypic traits	area side/top; convex hull area side/top; green area side/top; greener area side/top; height side/top; height above reference side; hue circular mean, std side/top; NIR mean, median, std side; projected shoot area; senescence index side/top; solidity side/top; width side/top	Numerical

The vector variable of phenotypic traits in Table 1 is now described [11–13]. Note that in most cases the view can be either side (S) or top (T).

- **Area top ($\mathbf{Area_{Top}}$):** Projected area of the plant from top view, measured in cm^2.

$$\text{Area}_{\text{Top}} = \text{Projected Area}_{\text{Top0}^\circ}$$

- **Area side ($\mathbf{Area_{Side}}$):** Mean projected area of the plant from 2 orthogonal side views (side 0°, side 90°), measured in cm^2.

$$\text{Area}_{\text{Side}} = \frac{\text{Projected Area}_{\text{Side0}^\circ} + \text{Projected Area}_{\text{Side90}^\circ}}{2}$$

- **Convex Hull Area side/top ($\mathbf{CHA_{view}}$):** Area of the convex hull containing the projected plant area from a specific view (for 'view'='Side' it is the average of the two sides), measured in cm^2.

$$\text{CHA}_{\text{view}} = \frac{1}{2}\left|\sum_{i=1}^{N-1}(x_i y_{i+1} - y_i x_{i+1}) + (x_N y_1 - y_N x_1)\right|$$

  Where (x_i, y_i) are the coordinates of the vertices of the convex hull.
- **Green Area side/top ($\mathbf{GA_{view}}$):** Projected green area of the plant from a specific view (Side or Top, for Side it is the average of the two sides), measured in cm^2.

$$\text{Green Area}_{\text{view}} = \sum_{\text{pixels}} H(60^\circ < \text{Hue} < 180^\circ)$$

  Where H is an indicator function that equals 1 if the condition is met and 0 otherwise.
- **Greener Area side/top ($\mathbf{GGA_{view}}$):** Projected greener area of the plant from a specific view (Side or Top, for Side it is the average of the two sides), measured in cm^2.

$$\text{Greener Area}_{\text{view}} = \sum_{\text{pixels}} H(80^\circ < \text{Hue} < 180^\circ)$$

  Where H is an indicator function that equals 1 if the condition is met and 0 otherwise.

- **Height side/top ($\mathbf{Height_{view}}$):** Height of the plant from a specific view (Side or Top, for Side it is the average of the two sides), measured in cm.

$$\text{Height}_{\text{view}} = \text{Height}_{\text{plant, view}}$$

- **Height Above Reference side ($\mathbf{Height_{ref}}$):** Height above a reference point, such as the base of the plant or soil surface, measured in cm.

$$\text{Height}_{\text{ref}} = \text{Height}_{\text{plant}} - \text{Height}_{\text{reference}}$$

- **Hue Circular Mean, Std side/top ($\mathbf{Hue\ Mean_{view}}$, $\mathbf{Hue\ Std_{view}}$):** Mean and standard deviation of the hue angle of the plant pixels from a specific view (side or top).

$$\text{Hue Mean}_{\text{view}} = \frac{1}{N}\sum_{i=1}^{N}\theta_i$$

$$\text{Hue Std}_{\text{view}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_i - \mu)^2}$$

  Where θ_i represents the hue angle of the i-th pixel, and μ is the mean hue angle.
- **NIR mean, median, std side ($\mathbf{NIR\ Mean_S}$, $\mathbf{NIR\ Median_S}$, $\mathbf{NIR\ Std_S}$):** Mean, median, and standard deviation of the Near Infrared (NIR) pixel intensity from a side view.

$$\text{NIR Mean}_{\text{S}} = \frac{1}{N}\sum_{i=1}^{N} I_i$$

$$\text{NIR Median}_{\text{S}} = \text{Median}(\{I_1, I_2, \ldots, I_N\})$$

$$\text{NIR Std}_{\text{S}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^2}$$

  Where I_i represents the NIR intensity of the i-th pixel, and μ is the mean NIR intensity.
- **Projected Shoot Area (PSA):** Sum of the projected plant area from three orthogonal images (side 0°, side 90°, top 0°), measured in cm^2.

$$\text{PSA} = \text{A}_{\text{Side0°}} + \text{A}_{\text{Side90°}} + \text{A}_{\text{Top}}$$

- **Senescence Index side/top ($\mathbf{Senescence\ Index_{view}}$):** Index of senescence calculated from green and greener areas from a specific view (side or top).

$$\text{Senescence Index}_{\text{view}} = \frac{\text{GA}_{\text{view}} - \text{GGA}_{\text{view}}}{\text{GA}_{\text{view}}} \times 100$$

  Where GA is the green area and GGA is the greener area.
- **Solidity side/top ($\mathbf{Solidity_{view}}$):** Measure of the compactness of an object, defined as the ratio of the area of the object to the area of its convex hull.

$$\text{Solidity}_{\text{view}} = \frac{\text{A}_{\text{Projected, view}}}{\text{CHA}_{\text{view}}}$$

- **Width side/top ($\mathbf{Width_{view}}$):** Width of the plant from a specific view (side or top), measured in cm.

$$\text{Width}_{\text{view}} = \text{Width}_{\text{plant, view}}$$

In Figure 1, the original images acquired from the LemnaTec platform are shown along with a representation of some of the main phenotypic traits investigated [14,15].

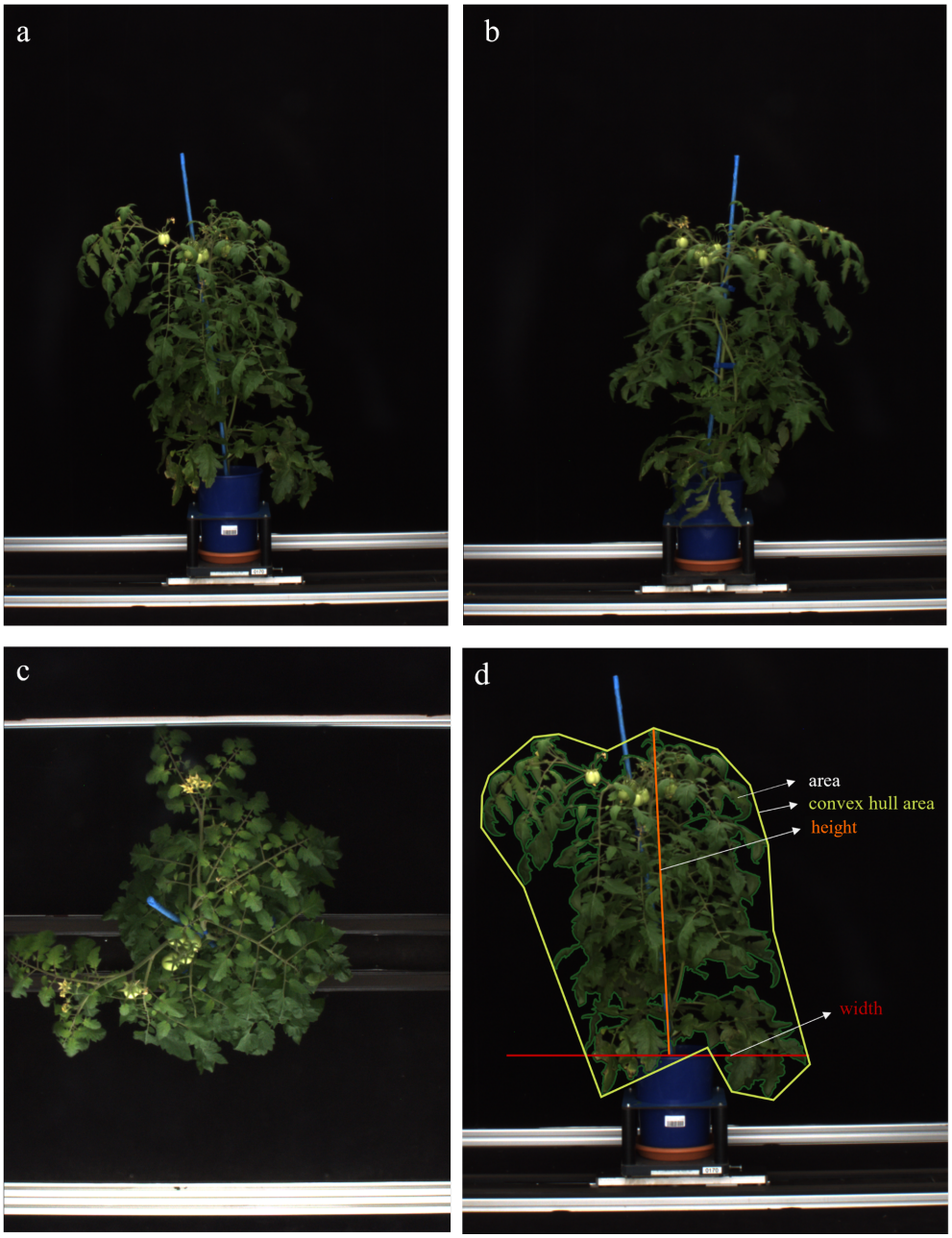

Figure 1. An instance of the photos that were taken and examined, in a: image captured from a side view (0°); in b: image captured from a side view (90°); in c: image captured from a top view; in d: a representation of some of the main phenotypic traits analyzed (area, convex hull area, height, width).

2.3. *Data Analysis*

R software, through the integrated programming environment RStudio (version 4.3.2; RStudio Team, 2024), and Python were used for statistical analysis and data visualization.

3. Data analysis

3.1. *Pre-processing*

The LemnaGrid software was used to extract tabular formatted data from the images, resulting in a CSV file. Next, a data cleaning step was performed by replacing the NaN (Not a Number) values with 0. This option was chosen as no other features had values as low as zero. The remaining values

were all positive, and they would have been eliminated by the Robust Scaler pre-processing. Indeed, pre-processing involved the adoption of a normalization technique [16]. We evaluated the performance of four available pre-processing techniques: the MinMax Scaler, Standard Scaler, L2-normalizing, and Robust Scaler. In our case, the Robust Scaler demonstrated the highest level of effectiveness and only L2-normalizing gave slightly different outcomes that were worth reporting for comparison between the two normalizers.

The data were subsequently transformed into time series, with each row representing a distinct observation of the same recorded plant at different points in time. The original dataset consisted of rows, each representing the characteristics of a plant at a particular time interval. The data has been arranged to depict each plant as a time series for every feature. Furthermore, to ensure consistent time intervals for analyzing all plants, a time binning procedure was conducted using a window size of 2 days. This was done because the duration of the experiment renders smaller time intervals irrelevant.

Image acquisition began when the plants were one month old and continued for two months, capturing a total of 16 distinct time points. At each time step, measurements were taken on 40 plants, with 5 replicates for each experimental condition. At the start of the experiment, an additional 40 plants were designated specifically for presetting. These presetting plants were excluded from the subsequent image analysis to avoid any potential bias in the experimental results. Instead, they underwent destructive biological analysis to provide baseline data necessary for calibrating and validating the experimental procedures. During the data cleaning phase, rows corresponding to these presetting plants were removed from the primary dataset to ensure that the analysis focused on the experimental subjects. However, the excluded rows were retained and used during the normalization process. Incorporating presetting data into normalization improves the scaling parameters' stability and robustness, ensuring that the normalization accurately reflects both experimental and presetting conditions. This method reduces potential variability while improving the consistency of feature scaling across the entire dataset.

3.2. *Statistical description*

Before applying the supervised machine learning techniques to the dataset, a statistical description was performed using PCA and K-means [17].

Principal Component Analysis (PCA) is used to reduce the dataset's dimensionality by identifying orthogonal directions, known as principal components (PCs), that account for the majority of the variance. Eigenvectors with higher eigenvalues represent directions with more variance in the data, while the principal components (PCs) are the covariance matrix's eigenvectors, sorted by eigenvalue (from highest to lowest). The first PC captures most of variance, then the second, and so on. After that, the data can be projected onto a lower-dimensional space spanned by the top k principal components, where $k < D$ and $D = 25$ is the number of features. The number of data points N were 5 replicas for each class ($5 \cdot 8 = 40$), each evaluated at 16 time stamps (from t_0 to t_{15}), which gives 640 total evaluations. After calculating the correlation matrix, any variables with a correlation greater than 0.7 were removed from this step of dimensionality reduction.

The partitioning of the data into unlabeled data points was also noteworthy because it allowed us to analyze where the data could naturally fall into the classes/thesis for which the experiment was designed. K-means clustering provides a straightforward and efficient method for data exploration and grouping. However, it requires prior specification of the number of clusters and is sensitive to centroids initial placement. In our case, we knew that the classes were either 2, 4, or 8. The K-means algorithm was implemented using the `K-means-rcpp` function in the RStudio programming environment. The analysis was carried out for each data acquisition time stamp (from t_0 to t_{15}). To ensure that the algorithm converged sufficiently, we performed a maximum of 200 iterations after the initial 100 iterations. In order to evaluate the effectiveness of clustering, we used a confusion matrix, which is a table that compares the class labels assigned by clustering with the true class labels. Additionally, we calculated the F-score, which is the harmonic mean of recall and accuracy. This measure indicates how well the final clusters align with the given class labels. As a result, selecting a cluster model with two groups was the best

choice for the data we were examining (k=2).

3.3. *Pipeline*

For the purpose of identifying any hidden patterns or predictors, a machine learning pipeline was developed. In the initial stage of the pipeline, a classifier was employed to account for supervised setting used to generate data. There are a number of classifiers that have been documented in the literature [16], and we have chosen a Naive-Bayes Classifier from among them. This decision was made because of the alignment between the classifier and the experiment, as predicted by the matching hypothesis. The Naive-Bayes Classifier was well-suited for the task due to the binary distinction between the classes, which is in line with the assumptions of the classifier and the binary experimental design.

To assess the impact of specific features on the performance of the chosen classifier, a non-parametric method known as *feature permutation importance* was used. This method involves determining the importance of each feature by calculating the increase in the model's prediction error after permuting the feature values. This method aids in determining which features are most important in the model's decision-making process.

3.4. *Feature Permutation Importance*

Let consider a dataset, $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, containing N data points. Each data point, $\mathbf{x}_i \in \mathbb{R}^d$, represents a vector of d features. The corresponding target variable is denoted by y_i. A trained model, $f : \mathbb{R}^d \to \hat{\mathbf{y}}$, predicts the target variable based on the features, where $\hat{\mathbf{y}} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N))$ represents the vector of predictions. For a given feature j ($1 \leq j \leq d$), it creates a permuted dataset, $\mathcal{D}^j = \{(\mathbf{x}_i^j, y_i)\}_{i=1}^N$, by randomly shuffling the values of feature j across all data points. The target variable vector, $\mathbf{y}$, remains unchanged. Then it evaluates the performance of the model, f, on the permuted dataset, $\mathcal{D}^j$, obtaining a new set of predictions, $\hat{\mathbf{y}}^j$. Consider the error function, E, as the performance metric:

$$E = L(f(\mathcal{D}), \mathbf{y})$$

where L is a loss function measuring the discrepancy between predictions and the target variable. The permutation importance score for feature j is calculated as the average increase in error caused by shuffling it:

$$\text{Importance}(j) = \frac{1}{N} \sum_{i=1}^{N} (E(f(\mathbf{x}_i, \mathcal{D}_{-i}^j)) - E(f(\mathbf{x}_i, \mathcal{D}_{-i})))$$

Here, $\mathcal{D}_{-i}^j$ denotes the permuted dataset excluding data point i, and $E(f(\mathbf{x}_i, \cdot))$ represents the error for a single data point.

A higher importance score for feature j indicates that changing its values significantly reduces the model's ability to predict the target variable. This implies that feature j is informative and adds significant value to the model's performance. The technique provides an intuitive understanding of feature importance by directly measuring the effect of rearranging a feature on model performance. Permutation importance is usually less computationally expensive than other feature importance methods, such as *feature elimination* or *coefficient analysis*. However, shuffling a single feature may have an indirect effect on the importance of other features due to potential correlations, and addressing this may necessitate appropriate correction for multiple comparisons. Additionally, with a large number of features, permutation importance can become computationally expensive due to repeated model evaluation.

The next phase in the pipeline was to determine which three of the 25 features were the most important. Given the relevance scores collected in the previous step of the pipeline, we had to choose a metric to calculate the final score and rank the features. The l2-norm was chosen over other metrics due to its ability to normalize all components between 0 and 1. This increased variance and accounted for overall trend over time.

Ultimately, we graphed the patterns of the most significant characteristics for each thesis in order to analyze the variations both qualitatively and quantitatively. Our main focus was on stressed plants, as we aimed to observe the disparities between receivers and non-recipient, as well as mutants and non-mutants.

4. Results and Discussion

To provide a better understanding of the rationale used to generate the results, we determined which features are more indicative. This is an A/B test, a statistical hypothesis test that compares two plant variations to determine if there is a difference in priming, genetic traits and stress responses. In an A/B test, the null hypothesis claims there is no difference between the control and variants groups. Unfortunately, the amount of samples gathered is insufficient to generate significant results across all 8 scenarios, as we attempted to verify. Therefore, we focused on analyzing the thesis into pairs of two, ensuring that each group has almost 30 occurrences [18–20]. We first ran a statistical analysis using K-means clustering without knowing the labels, and then evaluated the cluster's quality by matching the groups with an F-score. This allowed us to assess how effectively the data contained information that could distinguish between subgroups. Finally, we selected the Naive-Bayes Classifier because of the binary distinction between classes.

4.1. *Statistical description*

In this first step, K-means will assist us in determining how well all of the features together can identify different types of subgroups. Instead, the PCA will assist us in providing a preliminary graphical overview of which features are more representative in each group. We will still require a supervised approach because the percentage statistics will be available for each timestep and cannot be easily compared across different times.

4.1.1. *K-means*

The K-means algorithm provided valuable insights into the dynamics and structure of the data related to genotype (wildtype and mutant, Figure 2), treatment 1 (primed or no primed, Figure 3), and treatment 2 (stress or no stress, Figure 4) over time. Across all conditions, from t_0 to t_3, the two clusters showed non-uniform group mixing. However, from the fourth observation period onwards, the clusters began to separate more distinctly.
The presented plots display data grouped into two clusters in a two-dimensional representation (k=2). Each point represents an individual from the dataset, with distinct colors for each cluster. The lines indicate the boundaries between the clusters, and the dispersion of the points within each cluster reflects their cohesion. A legend specifies the colors of the cluster 1 and the cluster 2.
For genotype and treatment 1, the algorithm struggles to distinguish between the two groups from t_0 to t_3, but from t_4 onwards, the algorithm only accurately classifies half of the cases. Even though there was a clear separation in clusters, the algorithm's F-score remained at 50% for the detection of wildtype/mutant and primed/no-primed conditions. The analysis of treatment 2 revealed a clear pattern, indicating that the algorithm is more sensitive to differences related to stress occurrence. From the fourth observation period onwards, the initial mixing between the stress and no-stress groups was eliminated, with an F-score of 100%, demonstrating perfect classification between the two groups.

4.1.2. *PCA*

Principal component analysis (PCA) was applied to examine genotype (wildtype and mutant), treatment 1 (primed or no primed), and treatment 2 (stress or no stress). It was possible to simplify the data and identify the main sources of variation over time. The results were made more understandable by dividing them into three phases: t_0, t_5 and t_{10}, and t_{15}, which corresponded to the maturation stages of the plants.

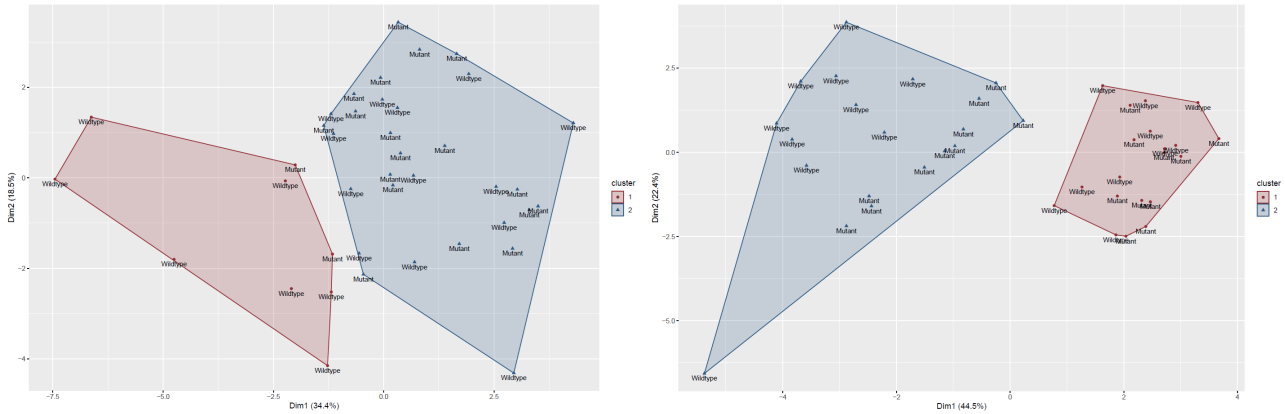

Figure 2. K-means classification for genotype (wildtype and mutant). The left plot illustrates the clustering of individuals from t_0 to t_3, wherein there is no clear distinction between wild-type and mutant plants. The plot on the right depicts the grouping of individuals from t_4 to t_{15}, revealing that the algorithm accurately classifies only half of the cases.

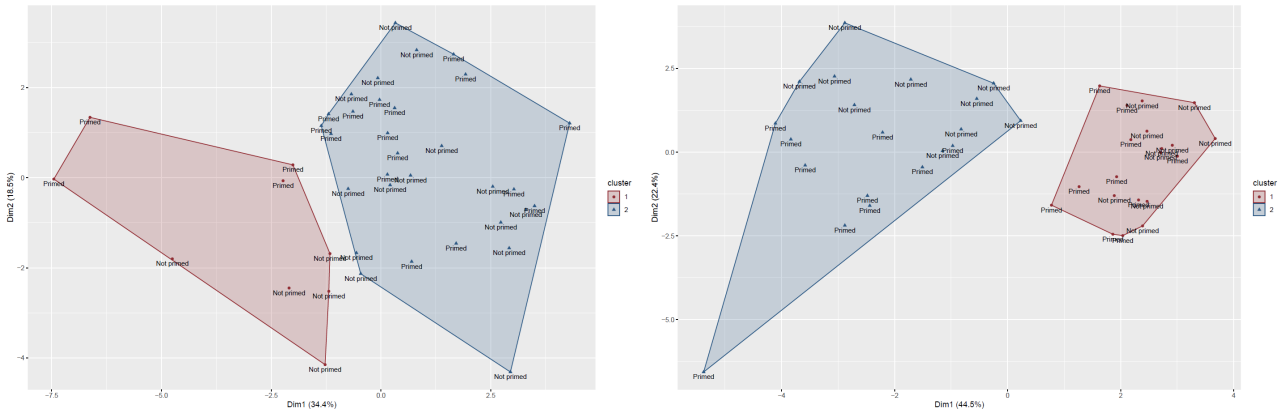

Figure 3. K-means classification for treatment1 (primed and not primed). The left plot illustrates the clustering of individuals from t_0 to t_3, wherein there is no clear distinction between primed and not primed plants. The plot on the right depicts the grouping of individuals from t_4 to t_{15}, revealing that the algorithm accurately classifies only half of the cases.

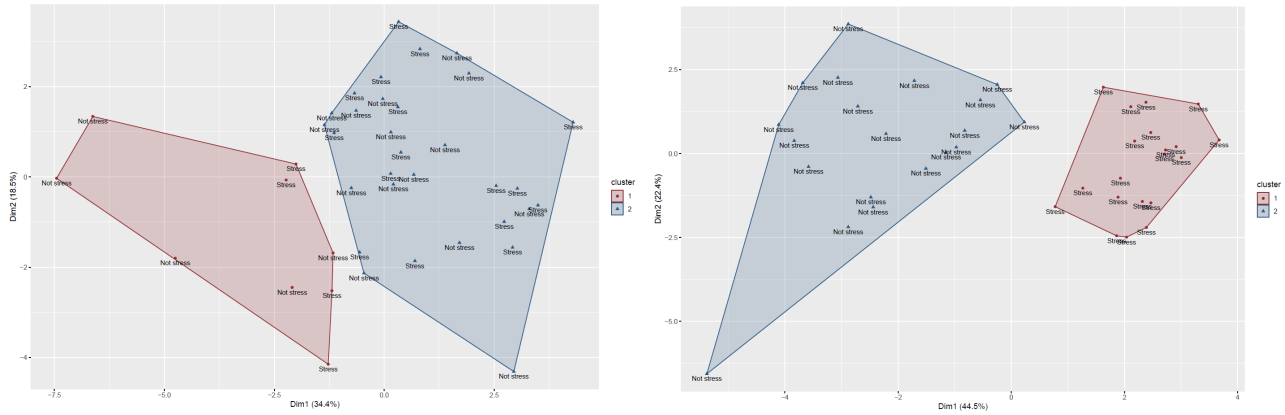

Figure 4. K-means classification for treatment2 (stress and not stress). The left plot illustrates the clustering of individuals from t_0 to t_3, wherein there is no clear distinction between stress and no stressed plants. The graph on the right depicts the grouping of individuals from t_4 to t_{15}, revealing that the algorithm accurately classifies 100% of the cases.

Analysis for genotype (Figure 5).

At time t_0, the primary sources of variation in the mutant group pertain to leaf color (hue): PC1 and PC2 account for 46.5% of the variance, with "hue circular mean top" contributing 15% and "hue circular mean side" contributing 25% as the primary variables. The wildtype group's plants exhibit a more ad-

vanced growth stage and reduced leaf water content. PC1 and PC2 explain 58.4% of the variance, with "senescence index top" (20%) and "NIR median side" (23%) as the main variables.
Within the middle interval (t_5 and t_{10}), both groups show no clear differentiation pattern. At t_5, the PCA of the mutant group reveals that PC1 and PC2 account for 70.7% of the variability, with "hue circular mean side" (25%) and "width top" (33%) as the primary variables; at t_{10}, PC1 and PC2 account for 65.7% of the variability, with "width top" (25%) and "NIR median side" (23%) as the primary variables. In wildtype group, at t_5 PC1 and PC2 explain 76% of the variability, with "width top" (22%) and "senescence index side" (24%) as the main variables; at t_{10}, PC1 and PC2 explain 70.6% of the variability, with "NIR median side" (28%) and "hue circular mean side' (28%) as the main variables.
At the end of the analysis (t_{15}), the mutant plants show more pronounced senescence than the wildtype group. The most significant factor appears to be the compactness of the plant structure. PC1 and PC2 explain 71% of the variance in mutant plants, with "senescence index top" (30%) and "NIR median side" (35%) as the main variables; for wildtype group, PC1 and PC2 explained 58.6% of the variance, with "NIR median side" (28%) and "solidity side" (33%) as the main variables.

Analysis for treatment 1 (Figure 6).
At t_0, irrespective of the group, the most significant variables are those that influence the color of the plant. For the primed group, PC1 and PC2 account for 58.9% of the variance, with "hue circular mean top" (23%) and "hue circular std top" (23%) as the primary variables. In the no primed group, PC1, and PC2 explain 54.2% of the variance, with "greener area top" (16%) and "hue circular mean top" (25%) as the main variables.
During the middle phase (t_5 and t_{10}), the PCA of the primed/no primed group reveals that the primary variables that contribute to the variance coincide with those that are typical of growth, suggesting that the senescence index is more prevalent in no-primed plants. For primed group, at t_5 PC1 and PC2 explain 70.6% of the variance and the main variables contributing to this variance are "width top" (22%) and "senescence index side" (36%); at t_{10} PC1 and PC2 explain 72.4% of the variance and the main variables here are "width top" (23%) and "solidity side" (35%). In no primed group, at t_5 PC1 and PC2 explain 68% of the variance, with "NIR median side" (30%) and "hue circular mean side" (38%) as the main variables; att_{10}, PC1 and PC2 explain 68% of the variance, with "width top" (25%) and "senescence index side" (43%) as the main variables.
In the final phase (t_{15}), it appears that plants subjected to priming showed stronger changes in color and height compared to no primed plants. The PCA of the primed group shows that PC1 and PC2 explain 61.7% of the variance, with "greener area top" (25%) and "width top" (34%) as the main variables; in no primed group, PC1 and PC2 explain 63.1% of the variance, with "circular mean top" (22%) and "NIR median side" (32%) as the main variables.

PCA analysis for treatment 2 (Figure 7)
At t_0, PCAs indicate that color-related variables carry the most weight in stressed plants. In the stress group PC1 and PC2 explain 54.4% of the variance, with "greener area top" (23%) and "hue circular mean side" (28%) as the main variables; in no stress group, PC1 and PC2 explain 57.4% of the variance, with "width top" (20%) and "senescence index top" (20%) as the main variables.
There is no particular distinction between the groups in the intermediate phases t_5 and t_{10}. In stress group, at t_5 PC1 and PC2 explain 53.8% of the variance, with "solidity side" (20%) and "width side" (18%) as the main variables; at t_{10} PC1 and PC2 explain 61.7% of the variance, with "senescence index top" (28%) and "hue circular std side" (26%) as the main variables. For no stress group, at t_5 PC1 and PC2 explain 58.7% of the variance, with "width top" (25%) and "senescence index side" (43%) as the main variables; at t_{10}, PC1 and PC2 explain 63.1% of the variance, with "NIR median side" (20%) and "solidity top" (32%) as the main variables.
In the final phase (t_{15}), the PCA suggests that stressed plants experience a significant change in the green area and plant height, while non-stressed plants have a strong connection to color. For stress group PC1 and PC2 explain 54.8% of the variance, with "greener area top" (25%) and "width top" (34%) as

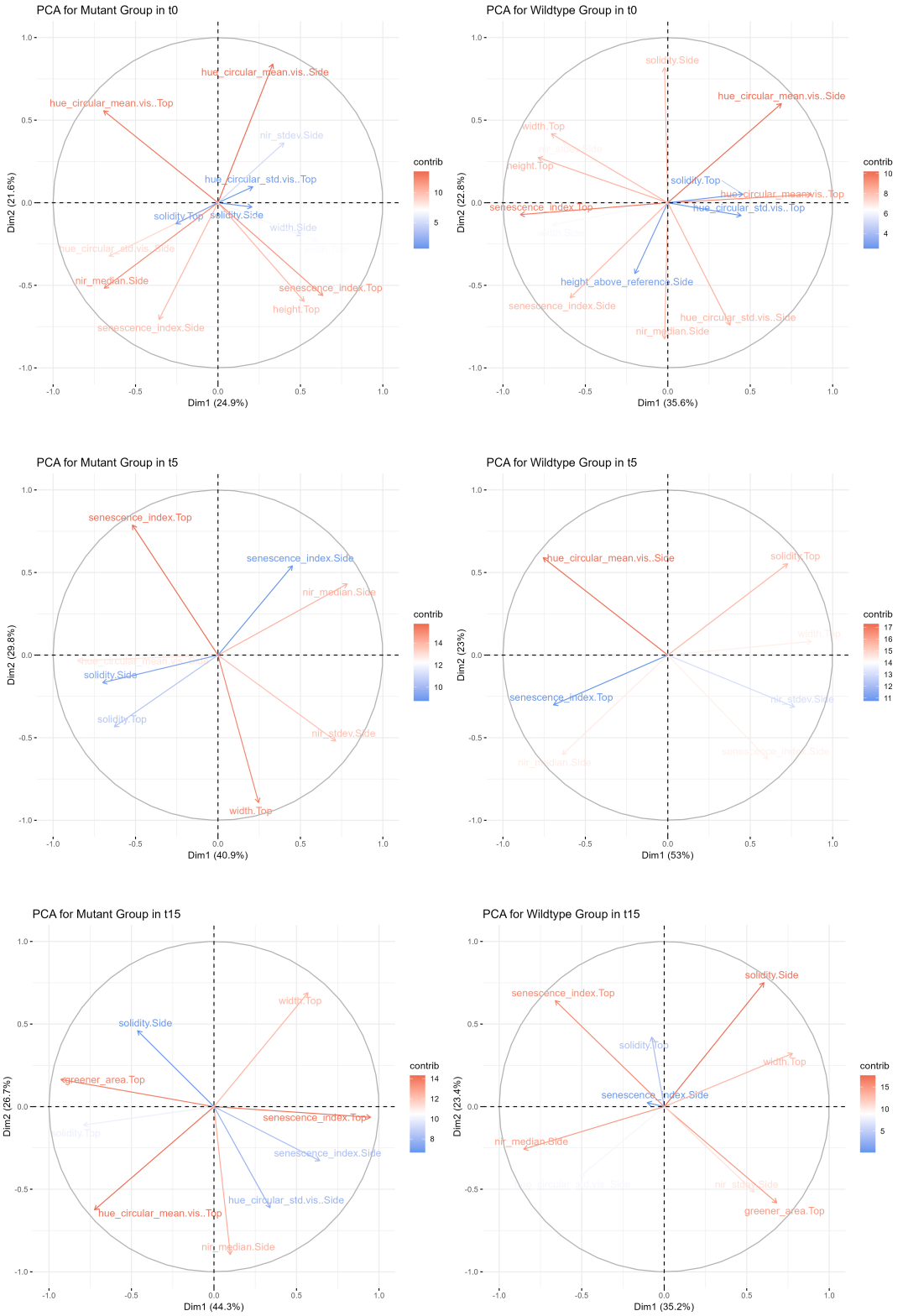

Figure 5. PCA for genotype (wildtype and mutant)

the main variables; for no-stress group, PC1 and PC2 explain 54.4% of the variance, with “hue circular mean top” (24%) and “NIR median side” (22%) as the main variables.

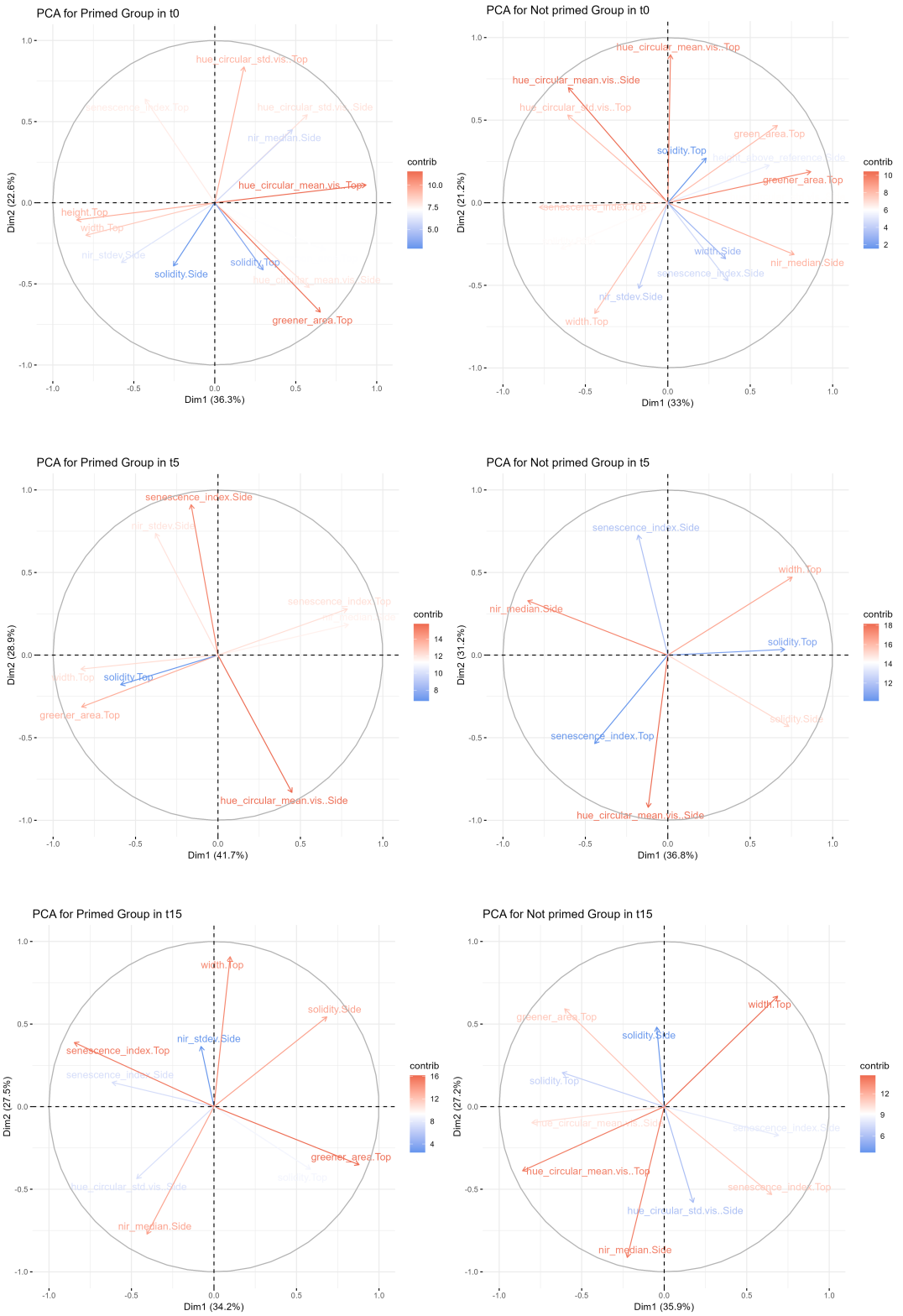

Figure 6. PCA for treatment1 (primed and no primed)

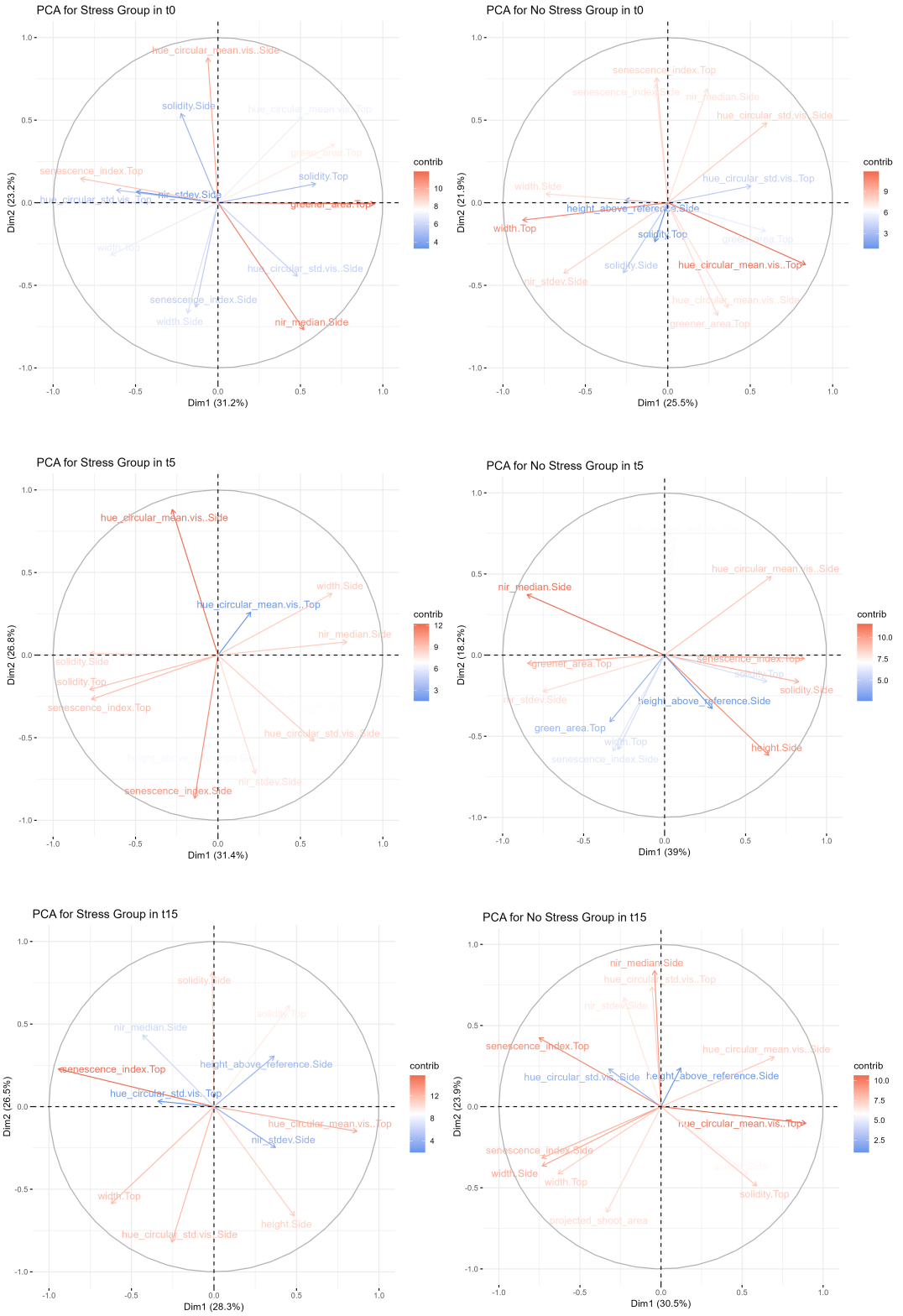

Figure 7. PCA for treatment2 (stress and not stress)

4.2. *Feature Permutation Importance through time*

We combined the machine learning pipeline results to determine the most important variables in the experiment. We identified three of the most important features for each classification and plotted

their values over time. We focused on stressed plants and highlighted the distinctions between classes and subclasses. As a result, interesting time intervals and patterns were revealed. Based on these new findings, more experiments should be carried out to determine a cause-and-effect relationship. The main results are presented in the following paragraphs.

The results of the Feature Permutation Importance Analysis over time obtained using Robust Normalization were useful in making several considerations. Regarding the classification of genotypes, as shown in Figure 8, the "senescence index top" (n.20) feature appears to be consistently relevant throughout the experiment. The features "hue circular mean side/top" (n.11-12) and "solidity top" (n.22) become increasingly important. Conversely, the features "area Top" (n.1), "convex hull area top" (n.3), "height side/top" (n.8-9), "hue circular std side/top" (n.13-14), "NIR median side" (n.16), "projected shoot area" (n.18), and "width side/top" (n.23-24) seem to be reducing in importance.

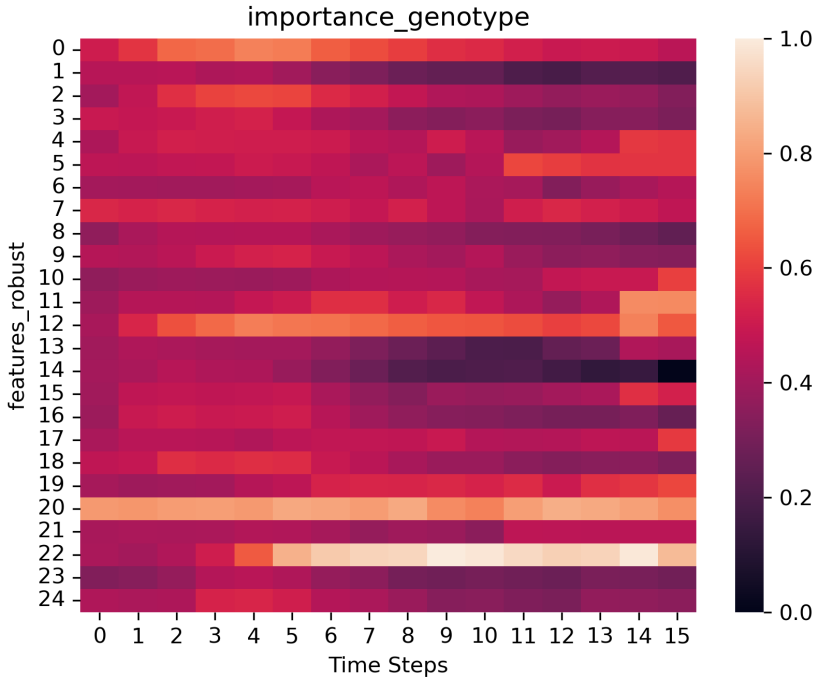

Figure 8. This heatmap obtained using robust scaled features shows the temporal dependency of the importance of features by permutation to classify the different genotypes (wildtype and mutant).

In the treatment 1 classification (Figure 9), it is important to note that the "green area side" feature (n.4) is the most significant, followed by "greener area side" (n.6), "hue circular mean side" (n.11), "senescence index top" (n.20), and "solidity side" (n.21) which gain importance over time. The features with the least impact are "area Top" (n.1), "convex hull area Side/top" (n.2-3), "hue circular std side" (n.13), "projected shoot area" (n.18), and "width side/top" (n.23-24). Additionally, the feature (n.14) "hue circular std top" is always equal to 0.

In treatment 2 classification (Figure 10), it is worth noting that most of the features have low importance, with "green area Top" (n.5), "height top" (n.9), "hue circular std side" (n.13), "NIR st dev side" (n.17), and "width top" (n.24) maintaining consistently low importance. "Hue circular std top" (n.14) and "width side" (n.23) almost always have a value of 0. The features "area side/top" (n.0-1), "convex hull area top" (n.3), and "projected shoot area" (n.18) are the variables that show the highest importance over time.

The first three features were extrapolated and analyzed from the Feature Permutation Importance Analysis using L2 normalization.

The three most significant genotype variables are "senescence index top", "solidity top" and "hue

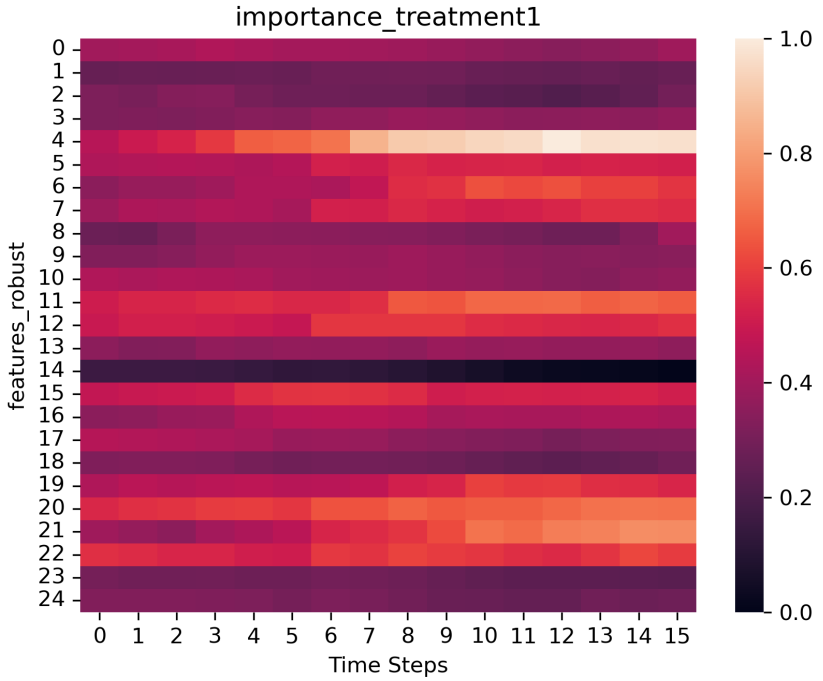

Figure 9. This heatmap obtained using robust scaled features shows the temporal dependency of the importance of features by permutation to classify if plants are primed or not primed.

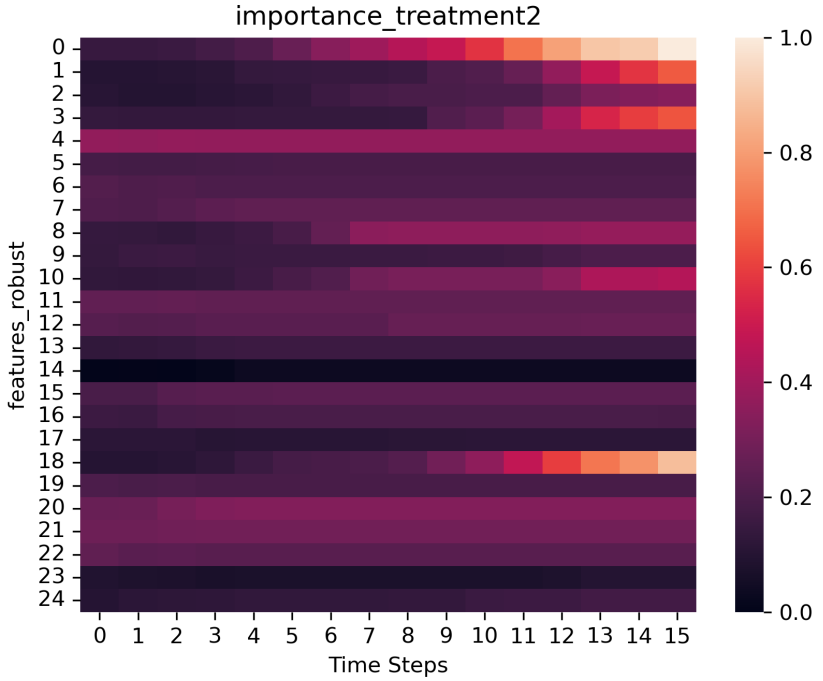

Figure 10. This heatmap, created using robust scaled features, shows temporal dependency of the importance of features by permutation to classify the stressed and not stressed plants.

circular mean top". Figure 11 shows a difference in "hue circular mean": in wildtype it starts at lower values and reaches lower values in comparison to mutant. In the wildtype, the "solidity" and "senescence index top" variables exhibit greater overlap during the central phase than in the mutant, which displays

a lower peak in “senescence index top” than the wildtype.

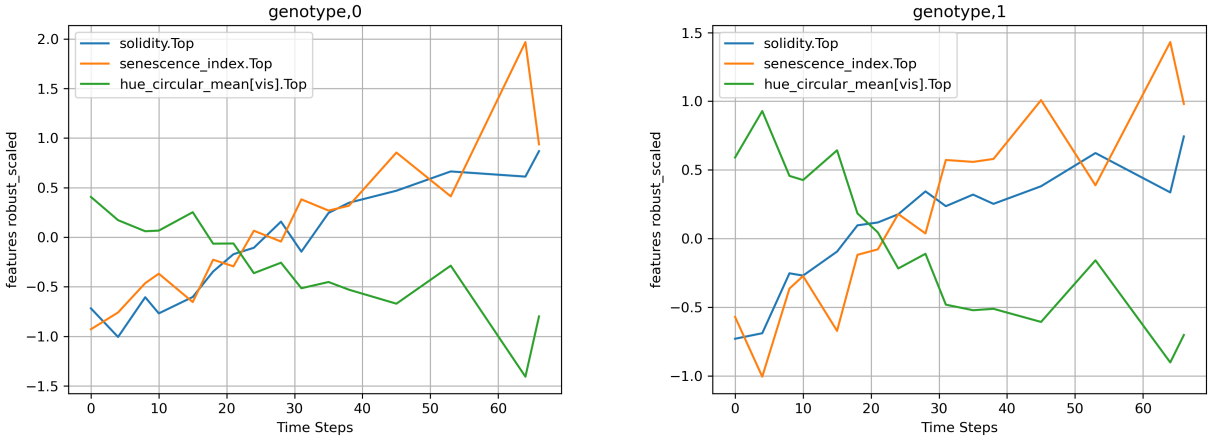

Figure 11. These 2 plots show the top3 features that distinguish the normal strain from the mutants over time (in L2-norm for importance).

For treatment 1 the three most important variables correspond to: “green area side”, “senescence index top” and “hue circular mean top”. The plots in Figure 12 show that the variables have similar trends in both groups but with some differences: (i) the greener area starts at lower values and remains lower over time for the no-primed plants, while in the primed ones, the values are higher and undergo more pronounced variations in the 10-40 time-step range; (ii) the senescence index top has a less stable trend in the 10-40 time-step range, with a slowdown in the increase in values for primed plants; (iii) the “hue circular mean” starts with lower values in the no-primed group compared to primed.

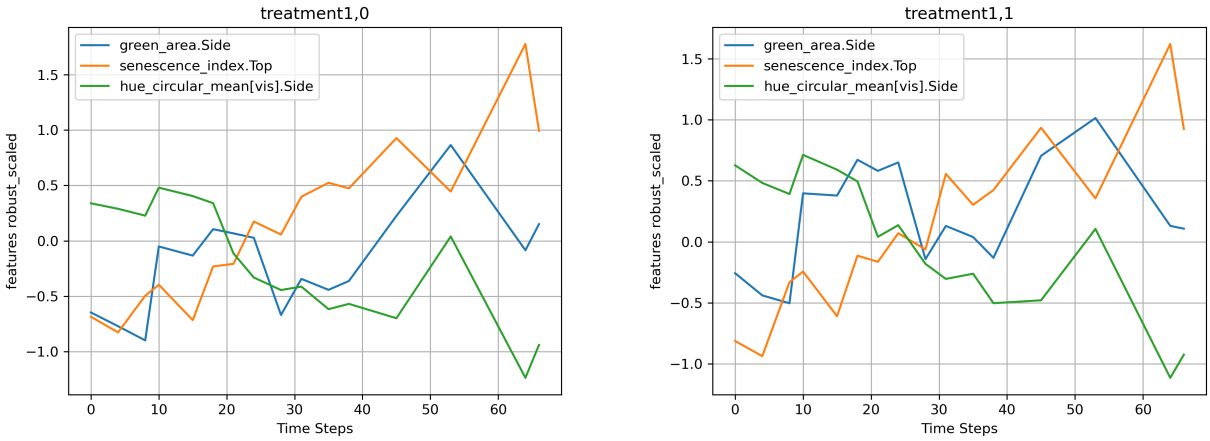

Figure 12. These 2 plots illustrate the top3 features (in L2-norm for importance) used to distinguish primed from not-primed plants over time.

The three most important variables for treatment 2 are “green area side’, “senescence index top” and “hue circular mean top”. In this dataset, there is a significant difference between the absence and presence of stress. Specifically, (i) the “side area” and “projected shoot area” variables exhibit an exponentially increasing trend over time before plateauing in the absence of stress. However, in the presence of stress, this exponential trend is not visible, and the variables’ values remain significantly lower; (ii) the “green area side” variable tends to increase in the absence of stress, although it shows a marked depression between 25 and 40 days, followed by a further increase and a sudden decrease. For plants stressed, however, the values of this variable tend to oscillate between highs and lows from 0 to 35 days, followed by a rapid increase to a maximum peak, and then a subsequent decrease and rise again.

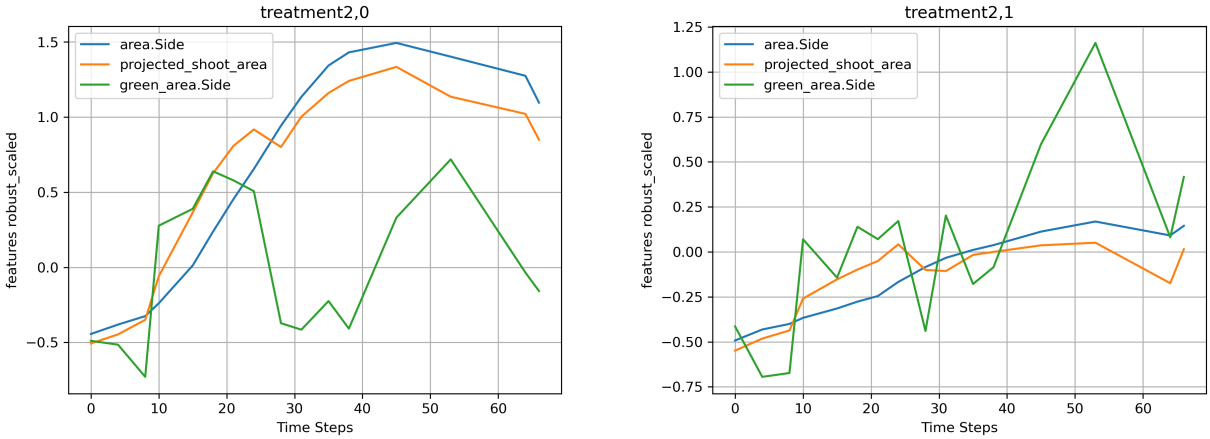

Figure 13. These two plots illustrate the temporal evolution of the top3 features (in L2-norm for importance) in differentiating between stressed and non-stressed plants.

With L2 normalization, it is possible to extrapolate and track the trend over time of the three most important characteristics for "treatment1" in stressed plants (Figure 12 and 13).

For primed/not primed plants (Figure 14), it is possible to see that : (i) "green area" and "hue circular mean" have initially greater importance for plants not primed, with a comparable trend later on; (ii) in primed plants, there is an oscillatory trend for the "senescence index" with a central peak occurring chronologically 10 days later than in non-primed plants. Eventually, the highest value is reached by the primed plants. In terms of genotype (Figure 15), the two genotypes demonstrate a similar pattern in "green area" and "hue circular mean" throughout the analysis, except for a higher peak observed at the end for the wildtype of "green area side"; the "senescence index" is more prominent in mutant plants, with an initial peak occurring 10 days earlier than in the wildtype, and a higher final value being attained.

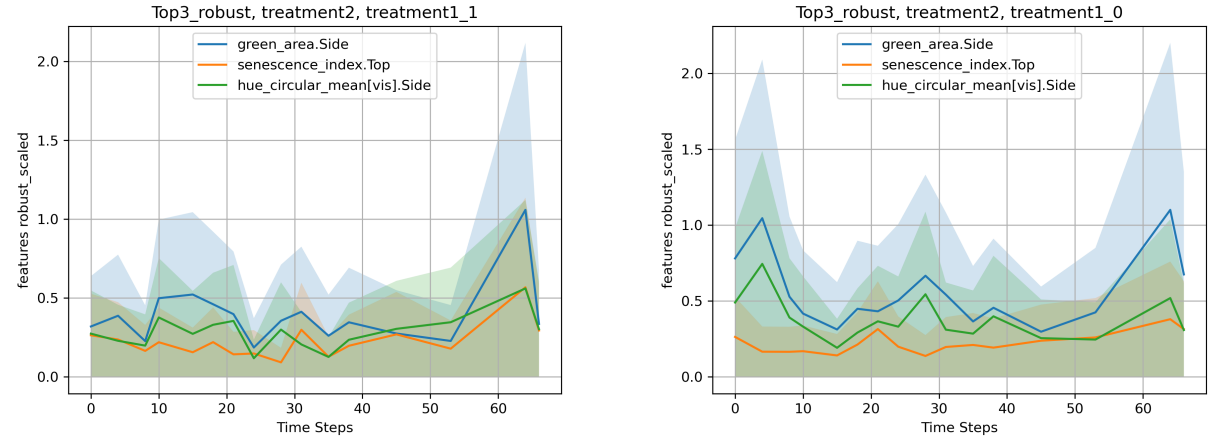

Figure 14. These 2 plots illustrate the temporal changes in the top3 features (in L2-norm for importance) between stressed plants that are primed and those that are not primed.

The difference between the stressed primed/no primed plants (in Figure 16) identified using Robust normalization is initially notable for "green area" and "hue circular mean", while in the end, it is strong only for "green area"; for "senescence index" the difference remains low except for the two peaks visible at time 20 and after time 60.

L2 normalization helps to visualize the differences between the two genotypes (Figure 16), especially

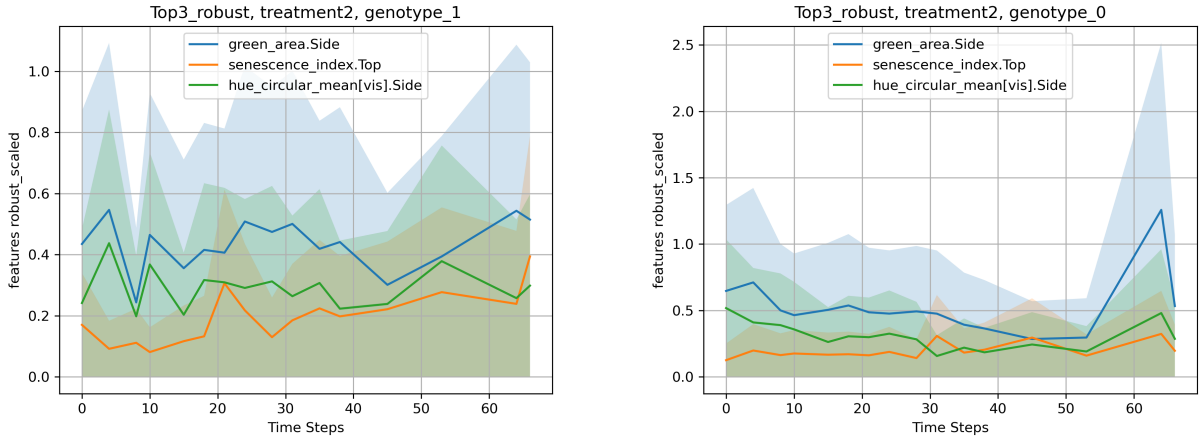

Figure 15. These two plots show through time the difference between the top3 features. The first two are between stressed mutant and wildtype plants and the last between stressed primed or not primed plants. The absolute value is used in the L2-norm for the first, and then in the semilog axis for the last two.

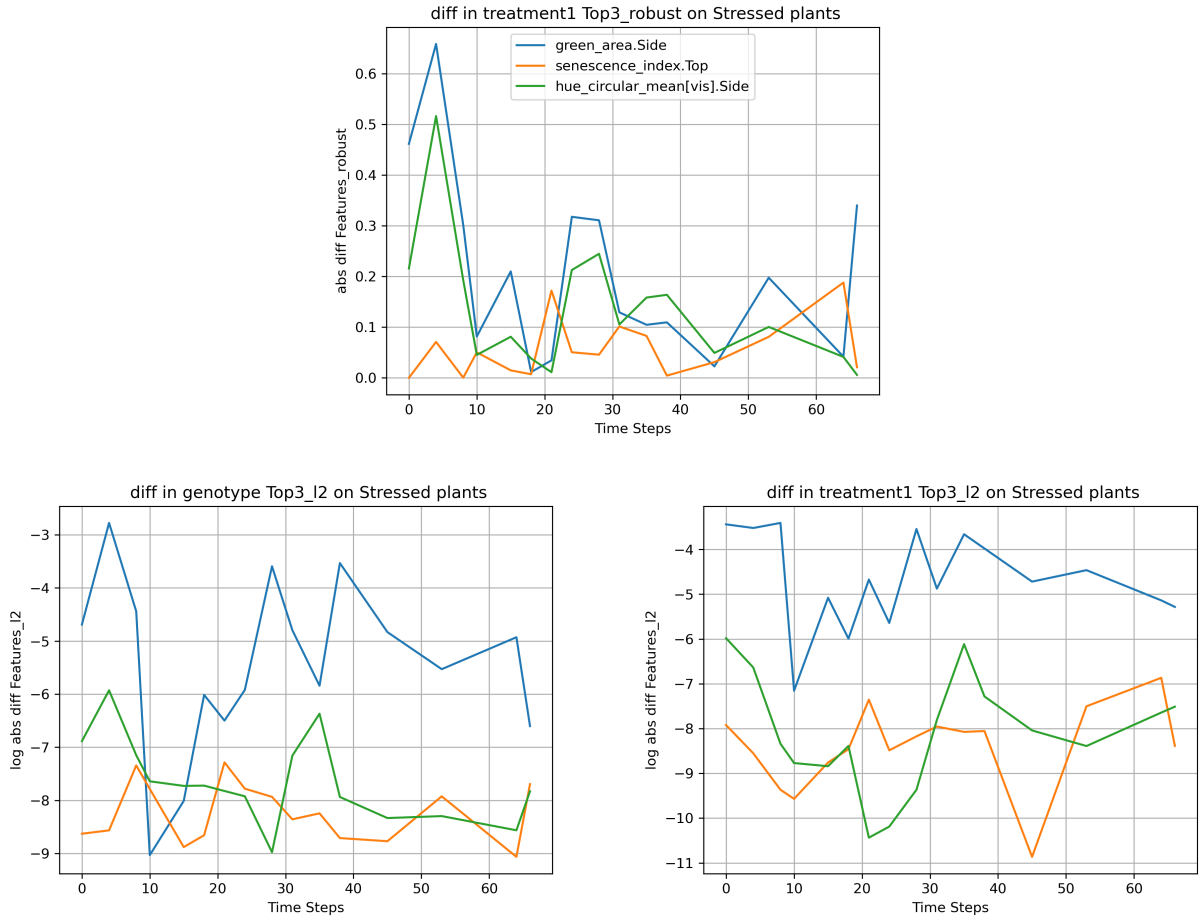

Figure 16. These plots depict through time the difference of the top three features (in absolute value and semi-log representation) between stressed primed or not primed plants.

in “convex hull area side” for the entire study period. The variations regarding “height above reference” are more pronounced at the initial and central points of the time series. The differences in “senescence index top” are slightly larger at the beginning and decrease at the end.

Plant "convex hull area" varies between the two types of plants in the initial and central phase (Figure 16); in the "senescence index" greater differences are noted at time 20 and the end of the analysis. The most significant differences in the "height above reference side" occur at the beginning and between times 30 and 40.

5. Discussion and Conclusions

The study emphasizes the importance of using both unsupervised and supervised learning techniques to analyze complex biological data. The combination of K-means clustering and PCA provided a solid foundation for initial data exploration and dimensionality reduction, while the Naive-Bayes Classifier and Feature Permutation Importance provided more detailed insights into feature significance and model performance. In plant science, limited data necessitates the use of multiple statistical and mathematical techniques to identify the most important and least important features. The ability to analyze numerous thesis can be enhanced by analyzing subgroups in order to increase the number of experimental units and ensure adequate sampling.

The study provided new insights into plant phenotypic responses to various treatments and genotypes. K-means clustering and PCA were useful for identifying key features and trends over time. The Feature Permutation Importance method identified the most influential features, allowing for a deeper understanding of plant responses. This comprehensive analysis emphasizes the importance of advanced statistical and machine learning techniques in plant phenotyping research.

In terms of K-means analysis, the best fit for the analyzed data was a cluster model with two groups, highlighting the importance of further investigation to fully understand the data's underlying structure. Initially, the clustering algorithm struggled to distinguish between groups due to overlapping properties. However, upon further analysis, more distinct separations became noticeable, particularly in the treatment2 (stress vs. no stress) analysis, which achieved perfect classification with an F-score of 100%. This demonstrates a strong response to stress treatments, making it a crucial factor in phenotypic variation.

The results indicate that variations in genetics and the application of experimental treatments, such as priming and stress, influence plant phenotypes. The mutant genotype exhibited increased variability and a more significant degree of senescence compared to the wildtype, suggesting a correlation between genotype and mechanisms of stress response.

Principal Component Analysis (PCA) was used to reduce the data's dimensionality and identify the main sources of variance throughout the experiment. The study concentrated on key phenotypic traits like "hue circular mean", "senescence index", "NIR median", and "solidity," which differed significantly between genotypes and treatments. These characteristics were essential in distinguishing between the mutant and wildtype groups, as well as between primed and non-primed, stressed and unstressed plants.

Analysis of time-series data revealed that certain phenotypic traits follow distinct temporal patterns. The "senescence index top" increased in the mutant group over time, reaching a peak at t_{14}. These temporal trends are critical for understanding the evolution of stress responses and the efficacy of priming over time.

Feature permutation determined which features had the greatest impact on the model's predictions. Features such as "senescence index top", "solidity top", and "hue circular mean top" were found to be particularly important. The L2-norm metric was used to rank these features over time, giving an extensive overview of their impact at different time points.

The detected features provide useful information for optimizing agricultural practices. For example, by identifying specific phenotypic markers linked to plant stress responses, we can devise precise intervention strategies, such as adjusting irrigation schedules to prevent water stress. Furthermore, understanding plant communication via VOCs can help to design more resilient crop varieties that can withstand environmental stressors, resulting in increased yield and reduced losses.

Future research should aim to validate these findings through additional experiments and expand the analysis to include other plant species and experimental conditions. The incorporation of more sophisticated machine learning models and techniques, such as deep learning, could improve the predictive power

and robustness of the phenotypic analysis. The limited size of the dataset posed challenges in formulating a robust and dependable definition of outliers. We conducted our PCA analysis concentrating on cases with reduced variability to minimize the influence of potential outliers In our forthcoming projects, we intend to integrate more resilient statistical methodologies, such as robust PCA and bootstrapping techniques. Furthermore, the development of user-friendly plant phenotyping software tools would increase the accessibility of these advanced techniques to researchers and practitioners. Finally, advances in plant phenotyping will contribute to more efficient and sustainable agricultural practices.

References

1. R. Pieruschka and U. Schurr, Plant Phenotyping: Past, Present, and Future, *Plant Phenomics*, vol. 2019, p. 7507131, Mar. 2019.
2. L. Li, Q. Zhang, and D. Huang, A Review of Imaging Techniques for Plant Phenotyping, *Sensors*, vol. 14, pp. 20078–20111, Nov. 2014.
3. A. Langstroff, M. C. Heuermann, A. Stahl, and A. Junker, Opportunities and limits of controlled-environment plant phenotyping for climate response traits, *Theoretical and Applied Genetics*, vol. 135, pp. 1–16, Jan. 2022.
4. F. Loreto and S. D'Auria, How do plants sense volatiles sent by other plants?, *Trends in Plant Science*, vol. 27, pp. 29–38, Jan. 2022.
5. J. Midzi, D. W. Jeffery, U. Baumann, S. Rogiers, S. D. Tyerman, and V. Pagay, Stress-Induced Volatile Emissions and Signalling in Inter-Plant Communication, *Plants*, vol. 11, p. 2566, Jan. 2022.
6. M. Nascimben, M. Venturin, and L. Rimondini, Double-stage discretization approaches for biomarker-based bladder cancer survival modeling, *Communications in Applied and Industrial Mathematics*, vol. 12, pp. 29–47, Jan. 2021.
7. M. B. Pouyan and D. Kostka, Random forest based similarity learning for single cell RNA sequencing data, *Bioinformatics*, vol. 34, pp. i79–i88, July 2018.
8. A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, Permutation importance: a corrected feature importance measure, *Bioinformatics*, vol. 26, pp. 1340–1347, 04 2010.
9. A. Cardellicchio, F. Solimani, G. Dimauro, A. Petrozza, S. Summerer, F. Cellini, and V. Renò, Detection of tomato plant phenotyping traits using YOLOv5-based single stage detectors, *Computers and Electronics in Agriculture*, vol. 207, p. 107757, Apr. 2023.
10. S. Kolhar and J. Jagtap, Plant trait estimation and classification studies in plant phenotyping using machine vision - A review, *Information Processing in Agriculture*, vol. 10, pp. 114–135, Mar. 2023.
11. J. Casadesús, Y. Kaya, J. Bort, M. M. Nachit, J. L. Araus, S. Amor, G. Ferrazzano, F. Maalouf, M. Maccaferri, V. Martos, H. Ouabbou, and D. Villegas, Using vegetation indices derived from conventional digital cameras as selection criteria for wheat breeding in water-limited environments, *Annals of Applied Biology*, vol. 150, no. 2, pp. 227–236, 2007.
12. P. Hüther, N. Schandry, K. Jandrasits, I. Bezrukov, and C. Becker, ARADEEPOPSIS, an Automated Workflow for Top-View Plant Phenomics using Semantic Segmentation of Leaf States, *The Plant Cell*, vol. 32, pp. 3674–3688, Dec. 2020.
13. J. Schrader, P. Shi, D. L. Royer, D. J. Peppe, R. V. Gallagher, Y. Li, R. Wang, and I. J. Wright, Leaf size estimation based on leaf length, width and shape, *Annals of Botany*, vol. 128, pp. 395–406, Sept. 2021.
14. T. A. Enders, S. St. Dennis, J. Oakland, S. T. Callen, M. A. Gehan, N. D. Miller, E. P. Spalding, N. M. Springer, and C. D. Hirsch, Classifying cold-stress responses of inbred maize seedlings using RGB imaging, *Plant Direct*, vol. 3, no. 1, p. e00104, 2019.
15. F. Zhu, M. Saluja, J. S. Dharni, P. Paul, S. E. Sattler, P. Staswick, H. Walia, and H. Yu, PhenoImage: An open-source graphical user interface for plant image analysis, *The Plant Phenome Journal*, vol. 4, no. 1, p. e20015, 2021.
16. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cour-

napeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-learn: Machine Learning in Python, tech. rep., June 2018. arXiv:1201.0490 [cs] type: article.

17. D. Arthur and S. Vassilvitskii, k-means++: the advantages of careful seeding, in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, USA: Society for Industrial and Applied Mathematics, Jan. 2007.
18. K.-S. Chiang, C. H. Bock, I.-H. Lee, M. El Jarroudi, and P. Delfosse, Plant Disease Severity Assessment - How Rater Bias, Assessment Method, and Experimental Design Affect Hypothesis Testing and Resource Use Efficiency, *Phytopathology*, vol. 106, pp. 1451–1464, Dec. 2016.
19. J. Zhou, J. Lu, and A. Shallah, All about Sample-Size Calculations for A/B Testing: Novel Extensions & Practical Guide, in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3574–3583, New York, NY, USA: Association for Computing Machinery, Oct. 2023.
20. R. R. d. Souza, A. Cargnelutti Filho, M. Toebe, and K. C. Bittencourt, Sample size and genetic divergence: a principal component analysis for soybean traits, *European Journal of Agronomy*, vol. 149, p. 126903, Sept. 2023.