NHL STENDEN
computer vision & artificial intelligence

Early Detection of PVY and PLRV in Potato Crops Using Hyperspectral Imaging and Machine Learning: A Cross-Variety Approach

NHL Stenden Professorship in Computer Vision & Artificial Intelligence

Tessa Postma,
Supervisor: Willem Dijkstra

Abstract—Potato Virus Y (PVY) and Potato Leafroll Virus (PLRV) are two of the most damaging viral diseases in European potato production, with PVY capable of reducing yields by up to 80%. Visual field inspection is slow, labour-intensive, and unreliable for early detection, especially during the pre-symptomatic phase when physiological changes occur weeks before visible symptoms appear. This study presents a machine learning pipeline combining deep learning-based leaf segmentation with XGBoost classification for early, cross-variety detection of both viruses using field-acquired hyperspectral images from a Specim FX-17 camera (900–1700 nm, 224 spectral bands) collected across two growth stages (early growth stage: May, late growth stage: June) and nine potato varieties. Individual leaves were manually annotated in CVAT and used to train a Mask R-CNN model. To separate the contribution of the classification pipeline from the leaf segmentation step, manually annotated ground truth masks were used as leaf masks in all main classification experiments. In a separate deployment evaluation, these were replaced by Mask R-CNN predicted masks to assess how segmentation quality affects end-to-end performance. After flat-field correction, pixel-wise ℓ_2-normalisation, and Savitzky-Golay derivative filtering, a leaf-level feature ablation was conducted across twelve feature combinations to identify the most informative per-leaf representation. Spectral mean combined with advanced shape features (207 features) matched the performance of the full 822-feature set while being four times more compact. These leaf-level features were aggregated per plant into a plant profile capturing canopy-level statistics, which helped reduce leaf-level noise while retaining class-relevant information. A canopy size ablation showed that removing small leaves severely degraded early-stage virus detection, indicating that the full canopy, including newly sprouted leaves, must be retained. Visual inspection of mean derivative spectra identified two biologically relevant regions (around 1080–1150 nm for cell structure and 1330–1400 nm for water stress). LDA-guided band selection partially confirmed these regions but revealed that the majority of the Top 20 discriminative bands span a broader window from approximately 1140 to 1560 nm, suggesting that LDA identifies discriminative information beyond what is visible in mean spectra plots. A spectral resolution ablation confirmed that the Top 20 LDA-selected bands consistently outperformed the full 204-band spectrum, suggesting that the additional bands introduce noise rather than useful signal. A deployment evaluation using Mask R-CNN predicted masks instead of ground truth annotations showed a drop in classification performance, from 88.5% to 61.5% in the late growth stage (June) and from 76.2% to 57.1% in the early growth stage (May), suggesting that segmentation quality is the main challenge for real-world deployment. Models trained in one growth stage also failed to generalise to the other, suggesting that a single universal model is probably not sufficient for reliable field deployment and that growth stage-specific training is likely required. Using Leave-One-Variety-Out (LOVO) cross-validation, cross-variety accuracies of $54.2 \pm 15.1\%$ (F1: 0.431 ± 0.137) across 6 qualifying varieties in the early growth stage and $56.5 \pm 21.8\%$ (F1: 0.482 ± 0.244) across 8 qualifying varieties in the late growth stage were found. These results show how difficult it is to separate disease signals from the natural spectral variation between potato varieties.

Index Terms—Hyperspectral Imaging, Machine Learning, Potato Virus Y (PVY), Potato Leafroll Virus (PLRV), Plant Disease Detection, XGBoost, Mask R-CNN, Plant Profile, Leave-One-Variety-Out Cross-Validation, Temporal Concept Drift.

1 Introduction

Crop diseases are one of the biggest threats to global food security, causing significant yield losses and economic damage across agricultural systems worldwide [1]. Potatoes are one of the world's most important staple crops, yet they are highly vulnerable to viral infections. Potato Virus Y (PVY) and Potato Leafroll Virus (PLRV) are two of the most damaging pathogens in European potato production. PVY alone can reduce yields by up to 80%, resulting in hundreds of millions of euros in annual losses across the European Union [2]. PVY causes a range of symptoms, including mosaic discolouration, leaf distortion, and vein or stem necrosis depending on the infecting strain. PLRV disrupts the plant's phloem transport system, causing characteristic leaf rolling and stunting. Both viruses spread rapidly through aphid transmission, and infected seed potatoes are a primary mechanism for continuing outbreaks across growing seasons [3]. Early and reliable detection of these viruses is therefore critical for effective crop management. Farmers who identify infected plants early can remove them before the virus spreads to neighbouring rows, substantially reducing the final economic impact. However, identifying PVY and PLRV during early growth stages through visual field inspection is slow, labour-intensive, and prone to expert error. The main challenge is that the earliest physiological changes often occur weeks before visible symptoms such as leaf curling, mosaic patterns, or yellowing appear [3]. This challenge is further complicated by the fact that the potato is cultivated across many distinct varieties, and this study covers nine of them, each with its own spectral baseline, meaning that detection methods trained on one variety may not reliably generalise to another. Conventional RGB imaging is restricted to three broad visible-range channels and therefore cannot resolve the subtle biochemical and structural changes in plant tissue that occur during early-stage infection [4]. Hyperspectral Imaging (HSI) addresses this limitation. By capturing reflectance across hundreds of narrow spectral bands simultaneously, HSI provides access to the complete spectral signature of plant tissue, revealing changes in pigment concentration, cellular

- *Tessa Postma is a Data Science for Life Sciences master student at the Hanze University of Applied Sciences.*
- *Willem Dijkstra is a researcher at the NHL Stenden Professorship in Computer Vision & Artificial Intelligence.*

water content, and internal leaf structure that are directly associated with viral infection [5]. The Specim FX-17 sensor covers 900–1700 nm, spanning the near-infrared (NIR, 900–1100 nm) and short-wave infrared (SWIR, 1100–1700 nm) regions using standard spectroscopy definitions [3]. Within this range, the NIR is primarily influenced by the leaf's internal cell structure [6], while the SWIR around 1350–1650 nm is particularly sensitive to leaf water content [7], making it relevant for detecting the water stress caused by viral disruption of vascular transport [3]. A previous study on this same dataset applied a Masked Spectral-Spatial Transformer (MaskedSST) for pixel-wise hyperspectral classification. That study concluded that pixel-level discrimination between PVY and PLRV was severely limited by spectral overlap across all 224 bands. Feeding a model the complete spectrum introduces the 'curse of dimensionality', causing overfitting to variety-specific or environmental noise rather than learning true disease-related spectral patterns [8]. This study takes a fundamentally different approach, combining deep learning-based leaf segmentation with classical machine learning classification. The central research question is: *How can hyperspectral imaging combined with machine learning be used to detect and classify Potato Virus Y and Potato Leafroll Virus at early infection stages across different potato varieties?* The main contributions are a leaf-level feature ablation study, a plant-level aggregation strategy (plant profile), and a Leave-One-Variety-Out cross-validation strategy to assess cross-variety generalisation. Three sub-questions guide the investigation:

- Which spectral bands or wavelength regions carry the most discriminative information for separating healthy, PVY-infected, and PLRV-infected potato plants, and how many bands are needed to maximise cross-variety accuracy?
- How effectively can a plant-level canopy feature vector support cross-variety virus classification under strict Leave-One-Variety-Out cross-validation?
- To what extent does model performance generalise across different growing stages, and what are the implications for practical field deployment?

2 State of the Art

This section reviews prior work relevant to the three core components of this study: hyperspectral disease detection, feature selection for spectral data, and gradient boosting classification.

2.1 Hyperspectral Disease Detection in Potatoes

HSI has been successfully applied to detect plant stress and disease across a range of crops. For potatoes specifically, Polder *et al.* used line-scan hyperspectral imaging combined with deep learning to detect PVY infections under field conditions, and showed that spectral signatures could reveal infection before visible symptoms appeared [9]. Gold *et al.* showed that hyperspectral measurements enable pre-symptomatic detection and differentiation of contrasting physiological effects of late blight and early blight in potatoes [10]. Despite these advances, much of the existing work is carried out in controlled greenhouse settings rather than in actual field conditions. While the dataset used in this study was also collected under controlled inoculation, meaning plants were knowingly infected before planting, the images were acquired in an open field, which introduces the kind of real-world spectral variation that greenhouse studies typically avoid. Generalisation across diverse genetic varieties has received limited attention. Griffel *et al.* evaluated spectral classification of PVY across multiple potato varieties and genotypes and found that variety-specific spectral baselines substantially complicate cross-variety detection [11]. Their findings highlighted that variety-specific spectral differences make cross-variety virus detection particularly difficult, which motivated the use of LDA-guided band selection in this study to identify wavelengths that are informative across varieties rather than specific to one.

2.2 Feature Selection and Dimensionality Reduction

Early HSI analysis relied heavily on classical machine learning approaches. Lapajne *et al.* compared multiple dimensionality reduction approaches for nematode detection in potato tubers, and found that targeted band selection outperformed full-spectrum models [8]. Nagasubramanian *et al.* used genetic algorithms with Support Vector Machines for hyperspectral band selection in soybean disease detection, this was used as a basis for the LDA-guided band selection in this study [12]. Spectral normalisation has been identified as a critical preprocessing step: Mazdeyasna *et al.* showed that ℓ_2-normalisation produces more robust spectral performance assessments than competing approaches [13]. The Savitzky-Golay derivative filter [14] is a well-established technique for extracting spectral slope information while suppressing high-frequency sensor noise [15].

2.3 Classification Methods

Gradient boosting methods, particularly XGBoost [16], have shown strong performance on tabular and engineered feature datasets in agricultural applications. Chen and Guestrin's original XGBoost formulation minimises a regularised objective combining a convex loss function and a complexity penalty, making it particularly suited to small, high-dimensional feature sets where overfitting risk is high [16]. George *et al.* confirmed that XGBoost-based ensembles achieve strong performance in plant disease classification tasks [17]. Instance segmentation through Mask R-CNN [18] has been widely used for individual leaf and plant part extraction in hyperspectral plant imaging, enabling clean spectral signatures to be obtained from individual leaves without contamination from background soil or overlapping plant material.

3 Materials and Methods

This section describes the dataset, preprocessing pipeline, feature engineering, and classification strategy.

3.1 Dataset and Sensor

The dataset consists of hyperspectral images of potato plants recorded under open-field conditions in the Netherlands. Images were acquired using a Specim FX-17 line-scan camera. To construct the hyperspectral data cube line by line, a motorized system moved the camera steadily over each plant. During acquisition, additional artificial illumination was used to provide consistent lighting and prevent signal gaps caused by the atmospheric water absorption window. The camera operates across the 900–1700 nm range, covering the near-infrared (NIR, 900–1100 nm) and short-wave infrared (SWIR, 1100–1700 nm) regions using standard spectroscopy definitions, across 224 spectral bands. Each hyperspectral data cube represents one potato plant or a small plant group, with spatial dimensions typically between 735 and 931 pixels in height and 640 pixels in width. A representative RGB preview generated from a data cube is shown in Figure 1. The dataset covers 9 distinct potato varieties: Agria, Arizona, Bartina, Challenger, Colomba, Fontane, Innovator, Seresta, and Spunta, from a total of 27 distinct field plots. Plants were recorded during two growth timepoints: early growth stage (May) and late growth stage (June), allowing temporal analysis of spectral changes associated with plant development. Three disease classes are represented: Healthy, PVY-infected, and PLRV-infected, though not all varieties carry all three classes. Colomba contains only Healthy plants, while Fontane, Innovator, and Seresta contain only Healthy and PVY-infected plants. Labels were assigned at the plant level, as the disease status of each plant was known prior to planting.

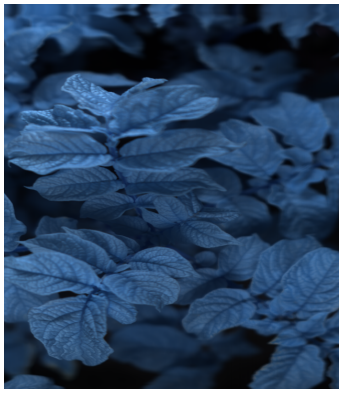


Fig. 1. Representative RGB preview of a potato plant from the dataset, generated from selected near-infrared bands of the hyperspectral cube.

3.2 Preprocessing

Flat-Field Correction. Raw FX-17 hyperspectral cubes contain sensor-level intensity values that must be converted to relative reflectance before spectral analysis. Flat-field correction is applied per-band using dark and white reference frames collected at the time of imaging:

$$R = \frac{I_{\text{raw}} - I_{\text{dark}}}{I_{\text{white}} - I_{\text{dark}}} \tag{1}$$

where I_{raw} is the raw pixel intensity, I_{dark} is the dark reference, and I_{white} is the white reference. For a subset of images without a white reference in the frame, a single white reference captured at the start of the acquisition was used. Negative values resulting from sensor noise are clipped to zero.

Spectral Band Cropping. The first and last 10 spectral bands were removed from each hyperspectral cube, reducing the usable spectrum from 224 to 204 bands (approximately 935–1646 nm). These edge bands are unreliable due to low signal-to-noise ratios at the spectral boundaries of the sensor, which flat-field correction cannot fully compensate for.

Spectral ℓ_2-Normalisation. Pixel-wise ℓ_2-normalisation removes brightness differences caused by illumination variation, leaf angle, and shadow, while preserving the spectral shape that encodes biochemical information. Pixels with zero intensity, such as those in complete shadow, are replaced by a neutral flat spectrum.

Savitzky-Golay Derivative. A first-derivative Savitzky-Golay filter (window of 15 bands, second-order polynomial) was applied to the normalised spectrum [14]. This converts the reflectance curve into a derivative spectrum that highlights the slopes and transition points of absorption features, where viral infections appear as amplitude changes. The derivative filter also suppresses slowly varying baseline offsets, further reducing sensitivity to illumination variation. Unless stated otherwise, all four preprocessing steps were applied to all experiments in this study. A preprocessing ablation confirming this choice is provided in Appendix A.

3.3 Manual Leaf Annotation and Mask R-CNN

Every individual leaf visible in each hyperspectral cube was annotated as a polygon instance mask in CVAT, using the Segment Anything Model (SAM) to generate initial polygon suggestions from an RGB preview generated from selected near-infrared bands, which were then manually reviewed and corrected where needed. Two mask types were created for each image: an instance segmentation mask assigning a unique colour-coded identifier to each individual leaf, and a class segmentation mask assigning the disease label colour to every leaf belonging to that plant (green for Healthy, red for PVY, yellow for PLRV). These CVAT annotations were used to train a Mask R-CNN model [18] for automatic leaf instance extraction. Once trained, the Mask R-CNN was applied to held-out images to automatically extract individual leaf instances. For the main classification experiments, ground truth annotations were used as leaf masks to isolate the performance of the classification pipeline itself. The Mask R-CNN predicted masks were evaluated separately as a deployment test (see Appendix E). The architecture is based on a ResNet-50 backbone, modified to accept 224-channel hyperspectral input instead of standard three-channel RGB images. Box and mask prediction heads were replaced with a two-class version (background and leaf) since disease classification is handled entirely by XGBoost. The model was trained for 250 epochs with a batch size of 2 using AdamW, with a base learning rate of 1e-4 and 5e-4 for the classification head, and no pretrained weights since the input channel count differs from standard RGB. The learning rate was halved when validation loss stopped improving for 3 consecutive epochs. Training augmentations included horizontal and vertical flips,

random 90-degree rotations, Gaussian noise, spectral band dropout, and intensity scaling. The plant-wise dataset split from Section 3.5 was used, with plant 1 for training and plant 2 for validation.

3.4 Classification Models

Three classifiers were evaluated in ascending order of complexity to justify the selection of a primary model. KNN ($k = 5$, Minkowski distance) served as a parameter-free baseline providing a lower-bound estimate of class separability. Random Forest (100 trees, Gini criterion, balanced class weights) provided an ensemble baseline with built-in feature importance estimates. XGBoost [16] served as the primary classifier, selected for its strong performance on tabular and engineered feature datasets and its built-in regularisation, which reduces overfitting on high-dimensional inputs [16, 17]. All three classifiers were first compared under identical baseline conditions. Hyperparameters for XGBoost were then tuned using GridSearchCV with Stratified 3-Fold cross-validation on the training set, testing 5,184 combinations across seven hyperparameters. The ranges and optimal parameters for both growth stage models are summarised in Table 1. Ranges were chosen based on standard XGBoost practice: shallow tree depths (max 6) and low subsample rates limit overfitting on the high-dimensional plant profile, gamma was capped at 0.5 since larger values tend to over-prune trees on small datasets, and learning rate was kept at or below 0.1 to allow sufficient boosting rounds without overshooting.

Table 1. XGBoost hyperparameter search ranges and optimal values for early and late growth stage models.

Hyperparameter	Search range	Early Growth	Late Growth
`colsample_bytree`	0.4, 0.6, 0.8	0.4	0.4
`gamma`	0, 0.1, 0.5	0	0
`learning_rate`	0.01, 0.05, 0.1	0.1	0.1
`max_depth`	2, 3, 4, 6	3	2
`min_child_weight`	1, 3, 5	1	1
`n_estimators`	50, 100, 200	200	200
`subsample`	0.6, 0.8, 1.0	0.8	0.6

All models used balanced class weighting to compensate for unequal class distributions. An additional $1.5\times$ penalty was applied to PVY sample weights to reduce false negatives, as PVY was consistently the most difficult class to detect. This multiplier was selected through a systematic grid search over $1.0\times$–$3.0\times$, where lower values underdetected PVY and higher values introduced false positives. A class weighting ablation confirming the impact of this choice on baseline performance is provided in Appendix B. The PVY penalty multiplier selection is confirmed in Appendix C.

3.5 Dataset Splitting

To prevent information leakage between training and evaluation sets, a strict plant-wise split was applied. Each physical field plot contains three individual potato plants, all of the same disease class (either all healthy, all PVY-infected, or all PLRV-infected), as disease status was assigned at the plot level prior to planting. Plant 1 from each plot was assigned to the training set, plant 2 to the validation set, and plant 3 to the held-out test set. This ensures no two plants from the same physical plot appear in both training and test, eliminating the risk of spectral leakage between closely located plants.

4 Experiments and Results

The experiments were conducted in the order in which methodological choices were locked in. Each experiment is described below. The baseline model comparison (Section 4.1) and feature ablation (Section 4.2) operate at the *leaf level*, evaluating individual leaf representations directly. All following experiments: spectral resolution ablation, canopy size ablation, temporal drift, and LOVO cross-validation operate at the *plant profile level*, using leaf-level features aggregated per plant as described in Section 4.2. Throughout this section, two evaluation strategies are used: a plant-wise split (plant 3 from each plot as the held-out test set) and Leave-One-Variety-Out (LOVO) cross-validation, in which all plants of one variety are withheld from training entirely and used as the test set to evaluate generalisation to unseen varieties. Where multiple folds are available (LOVO), results are reported as Mean $\pm$ SD across folds, all other experiments use a fixed train/test split and report single point estimates only.

4.1 Baseline Model Comparison

The three classifiers (KNN, Random Forest, and XGBoost) were compared under identical *leaf-level* baseline conditions: full spectral mean features (204 bands after edge-band removal), no LDA band selection, no size filtering, and no plant profile aggregation. This comparison serves to justify the choice of XGBoost as the primary classifier and to establish a reference point against which the gains from the full pipeline can be compared. Tables 2 and 3 present the results for both evaluation strategies.

Table 2. Baseline model comparison — LOVO cross-variety results (full spectrum, no plant profile aggregation, no LDA selection). Differences between models are small, no single classifier dominates across both growth stages.

Growth Stage	Model	Accuracy	Prec	Rec	F1
Early (May)	KNN	48.2%	0.403	0.354	0.344
Early (May)	Random Forest	49.0%	0.397	0.348	0.325
Early (May)	XGBoost	45.6%	0.432	0.363	**0.360**
Late (June)	KNN	40.5%	0.396	0.352	0.342
Late (June)	Random Forest	41.1%	0.434	0.347	0.317
Late (June)	XGBoost	42.7%	0.422	0.357	**0.337**

Table 3. Baseline model comparison — plant-wise split results (full spectrum, no plant profile aggregation, no LDA selection).

Growth Stage	Model	Accuracy	Prec	Rec	F1
Early (May)	KNN	51.2%	0.475	0.403	0.415
Early (May)	Random Forest	57.6%	0.562	0.454	0.468
Early (May)	**XGBoost**	**52.8%**	**0.517**	**0.480**	**0.486**
Late (June)	KNN	52.2%	0.525	0.513	0.515
Late (June)	Random Forest	59.5%	0.597	0.579	0.579
Late (June)	**XGBoost**	**62.4%**	**0.618**	**0.609**	**0.609**

Under plant-wise evaluation, XGBoost consistently achieved the highest F1-score in both growth stages (0.486 in May and 0.609 in June), confirming it as the strongest baseline classifier. Under LOVO cross-variety evaluation, differences

between the three models were small and no single classifier dominated across both growth stages, indicating that the cross-variety generalisation challenge is driven primarily by the data rather than the choice of model. XGBoost was therefore selected as the primary classifier for all following experiments.

4.2 Feature Ablation and Plant Profile Construction

Different feature types carry different amounts of noise and signal at the individual leaf level, so a systematic ablation was conducted over twelve leaf-level feature combinations to identify the most informative yet compact representation. Features were selected to cover three complementary aspects of leaf information: spectral texture (capturing the shape and variance of the reflectance curve), water-sensitive indices (targeting the known physiological effects of PVY and PLRV on vascular transport), and morphological shape (capturing structural changes to leaf geometry). Specifically, the features tested were: spectral mean (204 features), spectral standard deviation (204 features), 10th percentile spectrum (204 features), 90th percentile spectrum (204 features), Normalised Difference Water Index (NDWI, 1 feature), Moisture Stress Index (MSI, 1 feature), total leaf pixel count (1 feature), and advanced shape features including perimeter, eccentricity, and solidity (3 features). NDWI measures leaf water content by comparing reflectance at two near-infrared wavelengths [20], and MSI captures water stress by comparing short-wave infrared to near-infrared reflectance [21]. Each combination was evaluated on a held-out test set of 2,804 leaves, with XGBoost trained on 5,389 leaves using balanced class weighting, reporting macro F1-score, precision, recall, and overall accuracy, with the number of resulting features reported as a measure of dimensionality. The results are summarised in Table 4.

Table 4. Leaf-level feature combination ablation results ranked by macro F1-score. N = 2,804 test leaves.

Feature Combination	F1	Acc	Prec	Rec	Features
Everything (all features)	0.512	52.0%	0.555	0.528	822
All Textures (Mean+STD+P10+P90)	0.508	51.3%	0.549	0.526	816
Mean + Adv. Shape	**0.507**	**52.1%**	**0.545**	**0.524**	**207**
Mean + NDWI + MSI + Shape	0.501	51.8%	0.542	0.519	210
Mean + Size	0.500	51.6%	0.539	0.518	205
Mean + NDWI + MSI + Size	0.500	51.6%	0.538	0.517	207
Mean + Adv. Shape + Size	0.500	51.5%	0.540	0.518	208
Mean + NDWI + Size	0.499	51.5%	0.538	0.516	206
Spectral Mean Only	0.490	50.2%	0.537	0.513	204
Mean + STD	0.475	48.7%	0.517	0.496	408
Spectral STD Only	0.383	40.7%	0.435	0.419	204
Control: Morphology Only	0.251	34.6%	0.374	0.351	4

As shown in Table 4, the three top-performing combinations all achieved similar F1 scores (≈0.507–0.512), with precision and recall closely balanced across all three (precision ≈0.545–0.555, recall ≈0.524–0.528), indicating that the classifier is not simply favouring one class. Mean + Advanced Shape (207 features) matched the performance of the full 822-feature set while being four times more compact. Notably, adding spectral STD to the mean (Mean + STD, rank 10) reduced F1 to 0.475 with precision 0.517 and recall 0.496, relative to Mean Only (rank 9, F1 0.490, precision 0.537, recall 0.513), indicating that spectral variance at the individual leaf level introduces noise rather than signal. The morphology-only control (perimeter, eccentricity, solidity, and total pixel count — 4 features) confirmed that shape alone is not sufficient, achieving only F1 0.251 with precision 0.374 and recall 0.351, while combining spectral mean with shape features gave a consistent benefit over spectral mean alone. Spectral water indices including NDWI and MSI were also evaluated but did not improve performance beyond Mean + Advanced Shape, suggesting that the Savitzky-Golay derivative spectrum already captures the water-related absorption features these indices are designed to summarise, consistent with the preprocessing ablation in Appendix A.

A fundamental challenge in plant-level disease classification is that viral infection does not necessarily affect every leaf equally. Rather than classifying each leaf individually and voting, all leaves belonging to one physical plant are aggregated into a single plant profile feature vector before model training, preserving information about the distribution of spectral responses across the canopy. Based on these results, the Mean + Advanced Shape combination was selected as the leaf-level feature representation. These leaf-level features are then aggregated into a single plant profile feature vector as follows. In the first stage, per-leaf features are extracted for every individual leaf of a plant. In the second stage, these observations are aggregated across the entire canopy: the per-band mean of the spectral means across all leaves (204 dimensions), the mean of the advanced shape features (3 dimensions), the per-band standard deviation of the spectral means (204 dimensions), the standard deviation of the advanced shape features (3 dimensions), and two size features: total canopy pixel count and total leaf count. The complete plant profile is therefore a 416-dimensional vector per plant: 204 (mean spectra) + 3 (mean adv. shapes) + 204 (STD spectra) + 3 (STD adv. shapes) + 2 (size features). The per-band mean provides a stable spectral summary of the plant's condition, while the standard deviation across leaves captures canopy-level variation, which is informative since not all leaves are equally affected at the same time. All experiments from Section 4.4 onward use this plant profile level.

4.3 Exploratory Spectral Analysis and LDA Band Selection

Hyperspectral data contains 204 bands after edge-band removal, many of which are likely redundant or introduce noise. A two-step approach was used to identify the most informative spectral bands. First, mean derivative spectra were plotted per class and variety for both growth stages to visually identify regions with class separation. Two regions were selected for further investigation: Range A (bands 50–70, ≈1080–1150 nm), associated with leaf cell structure, and Range B (bands 120–140, ≈1330–1400 nm), associated with water stress. Second, Linear Discriminant Analysis (LDA) was applied to the 204-band mean spectra of each growth stage training set to statistically rank band discriminability across the three disease classes. LDA was chosen over Principal Component Analysis (PCA) because LDA is supervised and explicitly maximises class separation, making

it more appropriate for identifying bands that distinguish between healthy, PVY-infected, and PLRV-infected plants. PCA, being unsupervised, maximises variance without regard to class labels and may retain variance unrelated to disease status. Each band was scored by its average contribution across all LDA components.

Visual inspection of mean derivative spectra per class revealed consistent horizontal spectral shifts (x-axis shifts) between healthy and virus-infected classes. These shifts were most clearly visible in Range A (≈band indices 50–70, corresponding to 1080–1150 nm), where the valley positions of Healthy, PVY, and PLRV classes differ noticeably across most varieties and both dates (see Appendix D). Range B (≈band indices 120–140, corresponding to 1330–1400 nm) showed minimal visual class separation across most varieties, with the exception of Spunta in the early growth stage where a small x-shift was visible. This suggests that Range A is the most visually informative region for separating the three classes, while Range B shows little visible difference between them.

LDA partially confirmed these observations but also revealed structure beyond what is visible by eye. Several of the highest-scoring LDA bands fell within Range B in both growth stages (three bands in May, five in June), and one band at the edge of Range A (1140.4 nm) was selected in May. However, the majority of the Top 20 LDA-selected bands fell outside these two regions, covering a broader window from approximately 1140 to 1560 nm (see Figure 2). This indicates that discriminative spectral information extends further into the SWIR than visual inspection alone would suggest, and that LDA captures structure not visible in mean spectra. That LDA still selects bands from Range B, even though it shows little visual class separation, suggests that subtle amplitude differences in that region carry useful information that is not visible in the mean spectra plots. The majority of the selected bands fall within the 1140–1560 nm SWIR window. Bands around 1110–1130 nm reflect CH bond absorption in structural compounds such as cellulose and lignin [19], which break down under PVY-induced necrosis. Bands near 1450 nm capture OH bond absorption linked to water content [19], consistent with the water stress caused by PLRV disrupting phloem transport [3]. That LDA selects bands from both regions suggests the model captures both viral mechanisms.

4.4 Spectral Resolution Ablation Study

Eleven band configurations were evaluated on the plant-wise test set for both growth stages to determine how many and which bands are needed to maximise classification performance. All configurations use plant profile features, following the feature selection from Section 4.2. The configurations include the full 204-band spectrum, targeted subsets based on visually identified ranges or LDA-guided selection, and three additional baselines: PCA Top 20 components, every-10th band, and a random 20-band selection. The results are presented in Tables 5 and 6.

The Top 20 LDA-selected bands outperformed the full 204-band spectrum in both growth stages, confirming that the remaining bands introduce harmful noise rather than additional signal. Three additional baselines confirm that this gain is driven by supervised band selection rather than dimensionality reduction alone. Random 20 bands and every-10th band both performed worse than LDA Top 20 in both growth stages, showing that selecting any 20 bands is not sufficient. PCA Top 20 components, which reduce dimensionality unsupervised, performed notably worse than LDA Top 20 (F1 0.556 in May and 0.718 in June), confirming that retaining maximum variance without regard to class labels is less effective than explicitly maximising class separation, which is consistent with the motivation for choosing LDA over

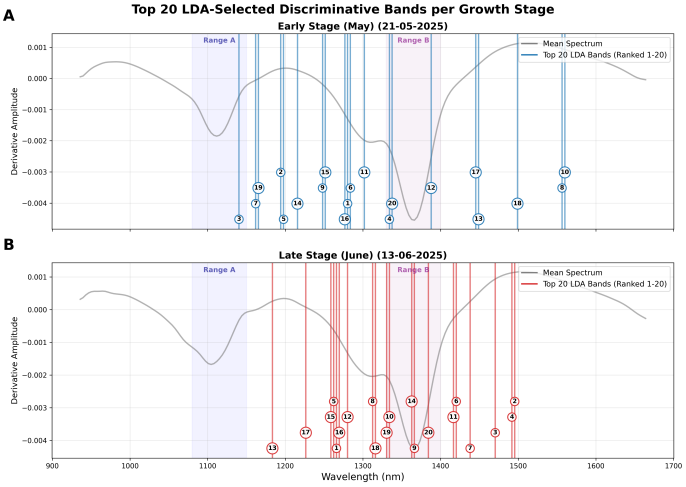


Fig. 2. Top 20 LDA-selected discriminative bands per growth stage overlaid on the mean first-derivative normalised reflectance spectrum. Panel A: Early Growth Stage (May, 21-05-2025). Panel B: Late Growth Stage (June, 13-06-2025). Range A (1080–1150 nm, blue) and Range B (1330–1400 nm, purple) are highlighted in both panels. The LDA-selected bands span a broader window from approximately 1140 to 1560 nm, covering more than the two visually identified ranges. Per-variety spectra for both dates are provided in Appendix D.

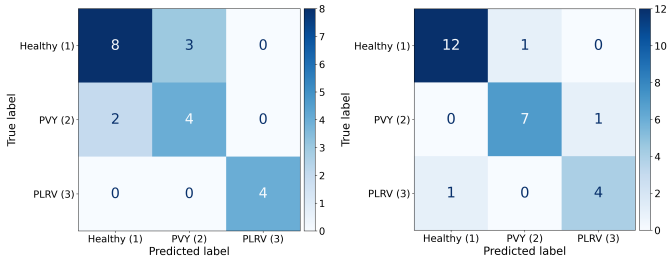
Table 5. Spectral Resolution Ablation — Early Growth Stage (May). All configurations use the fully optimised pipeline: tuned XGBoost hyperparameters (Table 1), plant profile aggregation, and 1.5× PVY penalty weight.

Band Configuration	F1	Prec	Rec	Acc	Features
LDA Top 20 Bands	**0.792**	**0.790**	**0.798**	**76.2%**	48
LDA Bands 11–20 (Middle 10)	0.751	0.739	0.773	76.2%	28
Full Hyperspectral (204 Bands)	0.742	0.742	0.742	71.4%	416
Random 20 Bands	0.709	0.707	0.715	71.4%	48
Every 10th Band (20 Bands)	0.707	0.713	0.715	71.4%	50
LDA Top 15 Bands	0.626	0.634	0.629	61.9%	38
Range A + B Combined	0.603	0.606	0.604	61.9%	92
PCA Top 20 Components	0.556	0.544	0.573	57.1%	28
LDA Top 10 Bands	0.543	0.543	0.545	57.1%	28
Range B Only (21 Bands)	0.483	0.467	0.518	52.4%	50
Range A Only (21 Bands)	0.440	0.449	0.437	52.4%	50

Table 6. Spectral Resolution Ablation — Late Growth Stage (June). All configurations use the fully optimised pipeline: tuned XGBoost hyperparameters (Table 1), plant profile aggregation, and 1.5× PVY penalty weight.

Band Configuration	F1	Prec	Rec	Acc	Features
LDA Top 20 Bands	**0.866**	**0.866**	**0.866**	**88.5%**	48
Range A + B Combined	0.768	0.765	0.774	80.8%	92
Random 20 Bands	0.768	0.765	0.774	80.8%	48
LDA Top 10 Bands	0.765	0.780	0.758	80.8%	28
LDA Bands 11–20 (Middle 10)	0.740	0.754	0.732	76.9%	28
Range B Only (21 Bands)	0.737	0.736	0.748	76.9%	50
Every 10th Band (20 Bands)	0.725	0.722	0.732	76.9%	50
PCA Top 20 Components	0.718	0.735	0.722	73.1%	28
LDA Top 15 Bands	0.704	0.737	0.700	76.9%	38
Full Hyperspectral (204 Bands)	0.662	0.701	0.674	69.2%	416
Range A Only (21 Bands)	0.527	0.525	0.531	61.5%	50

PCA described in Section 4.3. The ablation also shows the connection between the LDA-selected bands and the manually identified Range A and Range B: restricting the model to Range A and Range B combined performed notably worse than the Top 20 LDA bands (F1 0.603 vs. 0.792 in May; 0.768 vs. 0.866 in June), consistent with the finding that LDA identifies discriminative bands spread across a wider 1140–1560 nm window rather than confined to these two regions alone. An unexpected result was found when the Top 20 bands were split into two equal halves. In June (mature plants), the Top 10 outperformed the Middle 10 (80.8% vs. 76.9%), as expected from an importance ranking. In May (young plants), however, the Top 10 completely collapsed to 57.1%, while the Middle 10 achieved 76.2%. This suggests that in young plants, the Top 10 bands alone do not contain enough independent information to classify disease reliably, and that the Middle 10 bands add information that the Top 10 is missing. The full Top 20 is therefore needed to perform well across both growing stages. Confusion matrices for both growth stages under the optimal Top 20 LDA band configuration are shown in Figure 3. This configuration achieved the best plant-wise results of the entire study, with 76.2% accuracy (F1: 0.792, Recall: 0.798) in May and 88.5% accuracy (F1: 0.866, Recall: 0.866) in June, and the 1.5× PVY penalty can be seen to successfully balance PVY recall across both growth stages.



(a) Early Growth Stage (May) (b) Late Growth Stage (June)

Fig. 3. Confusion matrices for the optimal XGBoost models (Top 20 LDA bands, 1.5× PVY penalty, plant-wise test set). Panel A shows the early growth stage (May) and Panel B the late growth stage (June).

PVY misclassification is higher in May than June, consistent with its biology: in the early growth stage, PVY has spread through the plant but has not yet caused the cell damage and leaf discolouration that produce strong spectral differences in the 1140–1560 nm range where the LDA-selected bands are concentrated. The Specim FX-17 is primarily sensitive to internal leaf structure and water content, and in May these changes are still small and unevenly spread across the canopy, making classification harder. By June, the damage is more widespread and consistent across leaves, producing a stronger and clearer spectral signal for the classifier to detect. A per-class SHAP analysis (see Appendix G) confirms that PVY and PLRV rely on partially distinct spectral regions, consistent with their different biological mechanisms.

4.5 Canopy Size Ablation Study

To assess whether leaf size contributes signal or noise when constructing the plant profile, size filters were applied in both directions across a range of standard deviation thresholds relative to the per-date mean leaf pixel count. A lower-bound filter removes small leaves by keeping only leaves above mean $-$ factor $\times$ std, while an upper-bound filter removes large leaves by keeping only leaves below mean $+$ factor $\times$ std. Both filters were applied consistently across train, validation, and test sets before computing the plant profile. Table 7 presents the results for both growth stages.

Table 7. Canopy size ablation results. Lower-bound filters remove small leaves and upper-bound filters remove large leaves. XGBoost with plant-wise test set and fixed optimized parameters.

STD Factor	Leaves Kept	May Acc	May F1	June Acc	June F1
No filter (baseline)	**100%**	**76.2%**	**0.792**	**88.5%**	**0.866**
Lower-bound filters (removes small leaves)					
2.0 STD	∼98%	76.2%	0.792	88.5%	0.866
1.5 STD	∼93%	76.2%	0.792	88.5%	0.866
1.0 STD	∼84%	76.2%	0.792	88.5%	0.866
0.5 STD	∼69%	57.1%	0.512	80.8%	0.775
0.0 STD	∼50%	57.1%	0.520	57.7%	0.506
Upper-bound filters (removes large leaves)					
2.0 STD	∼98%	66.7%	0.674	84.6%	0.816
1.5 STD	∼93%	61.9%	0.544	76.9%	0.716
1.0 STD	∼84%	61.9%	0.556	84.6%	0.810
0.5 STD	∼69%	52.4%	0.466	80.8%	0.759
0.0 STD	∼50%	42.9%	0.354	92.3%	0.917

For the lower-bound filters, very mild filtering (2.0, 1.5, and 1.0 STD) had no impact on performance, indicating that the plant profile is already robust to extreme small-leaf outliers. Only when the filter became aggressive enough to remove a meaningful portion of the canopy (0.5 STD and below) did performance degrade, dropping from 76.2% to 57.1% in May and from 88.5% to 57.7% in June at the most aggressive threshold. These results confirm that small leaves carry useful signal rather than noise, consistent with PLRV's known disruption of phloem transport in young leaf tissue [3], suggesting that newly sprouted leaves may carry early-stage viral signatures that are absent or weaker in mature leaves. The upper-bound filters revealed an asymmetry between growth stages. In June, removing large leaves had relatively little impact at mild thresholds (upper 2.0 and 1.0 STD both achieved 84.6% accuracy), and the most aggressive upper-bound filter (upper 0.0 STD, keeping only below-mean leaves) actually improved performance above the no-filter baseline, reaching 92.3% accuracy (F1: 0.917). This suggests that in mature plants, very large leaves may introduce canopy-level noise that partially obscures disease signals. In May, however, removing large leaves consistently degraded performance, with the most aggressive filter dropping accuracy to 42.9% (F1: 0.354). Since no single filter consistently improves both growth stages simultaneously, the No Filter configuration was retained for all subsequent experiments.

4.6 Temporal Concept Drift

A temporal cross-validation experiment was conducted to assess how stable model performance is across plant development stages within the same growing season. The May model was evaluated on the June test set and vice versa, since May and June represent different growth stages rather than independent seasons. Performance degradation was severe in both directions, as shown in Table 8. The May model achieved

only 49.4% accuracy on June data, compared to 76.2% on the May test set. The June model dropped to 28.6% on May data, which is below random chance for a three-class problem. To understand why, the June test set was split into two equal halves based on total canopy pixel count, separating smaller plants (bottom 50%) from larger mature plants (top 50%). The May model scored 61.5% on small June plants but only 36.8% on large mature plants, suggesting that the classifier is partly relying on plant size to make its predictions. When plant sizes shift between growth stages, these learned patterns break down. This is an example of concept drift [22], a limitation that has received little attention in the existing HSI disease detection literature.

Table 8. Temporal concept drift evaluation results. Models trained on one growth stage are tested on the other, also split by plant size.

Training Stage	Test Stage	Plant Size Group	Accuracy	F1
Early (May)	Late (June)	Entire dataset	49.4%	0.220
Early (May)	Late (June)	Small plants (bottom 50%)	61.5%	0.254
Early (May)	Late (June)	Large plants (top 50%)	36.8%	0.179
Late (June)	Early (May)	Entire dataset	28.6%	0.190
Late (June)	Early (May)	Small plants (bottom 50%)	31.2%	0.198
Late (June)	Early (May)	Large plants (top 50%)	25.8%	0.205

4.7 Final Cross-Variety Generalisation (LOVO)

For the LOVO cross-validation experiments, all plants of one variety are withheld from training entirely and used exclusively as the test set, evaluating how well the model generalises to potato varieties it has never seen during training. Varieties with only a single disease class were excluded from LOVO evaluation, leaving 6 qualifying varieties in May (Agria, Arizona, Bartina, Challenger, Seresta, and Spunta) and 8 in June (Agria, Arizona, Bartina, Challenger, Fontane, Innovator, Seresta, and Spunta). The fully optimised pipeline (No Size Filter, Top 20 LDA bands, 1.5× PVY penalty, tuned XGBoost hyperparameters) was evaluated using Leave-One-Variety-Out cross-validation at the plant canopy level. Table 9 presents the final cross-variety results reported as Mean ± SD across varieties. Comparing these results to the baseline LOVO results in Table 2 reveals a clear improvement: from 45.6% to 54.2% accuracy in May and from 42.7% to 56.5% in June, confirming that the gains achieved by the full pipeline (LDA band selection, plant profile aggregation, and tuned hyperparameters) are real and not just due to the choice of classifier.

Table 9. Final LOVO cross-validation results across completely unseen potato varieties, reported as Mean ± SD across qualifying varieties (6 in May, 8 in June).

Dataset	Acc (Mean±SD)	Prec (Mean±SD)	Rec (Mean±SD)	F1 (Mean±SD)
Early (May)	$54.2 \pm 15.1\%$	0.462 ± 0.151	0.445 ± 0.147	0.431 ± 0.137
Late (June)	$56.5 \pm 21.8\%$	0.463 ± 0.271	0.533 ± 0.234	0.482 ± 0.244

For a three-class problem the random baseline is 33.3%. The pipeline achieved 54.2% in May and 56.5% in June on average across varieties, both above chance. The high standard deviation (15.1% in May, 21.8% in June) reflects that generalisation varies widely depending on the held-out variety. Cross-variety accuracy dropped substantially compared to the plant-wise evaluation: from 76.2% to 54.2% in May, and from 88.5% to 56.5% in June. This drop shows that natural spectral differences between potato varieties often hide the subtle spectral changes caused by early-stage viral infection. While the model successfully learned disease signatures within known varieties, generalisation to completely unseen varieties remains very difficult, consistent with the findings of Griffel *et al.* [11]. Per-variety results vary widely, with Spunta and Challenger generalising well in June (F1 0.908 and 0.850) while Bartina and Innovator performed poorly (F1 0.222 and 0.250), suggesting that some varieties share more spectral similarity with the training set than others (see Appendix F).

5 Discussion, Conclusion and Future Work

This section discusses the main findings of the study, addresses the central research question and sub-questions through its conclusions, and outlines directions for future work.

5.1 Discussion

The most important finding of this study is that using the full hyperspectral spectrum is not only unnecessary for cross-variety potato virus detection, but is actually harmful to classification performance. The Top 20 LDA-selected bands consistently outperformed the 204-band full-spectrum features in both growth stages, contradicting the common assumption that more spectral information leads to better classification. This result is in line with the findings of Lapajne *et al.* [8] in a multi-variety, open-field setting, and suggests that the hyperspectral curse of dimensionality is a real obstacle for variety-generalised disease detection [8, 12]. Three additional baselines all performed worse than LDA Top 20, confirming that the gain is driven by the specific discriminative bands LDA identifies rather than dimensionality reduction alone. It should also be noted that two optimal hyperparameter values landed at grid boundaries: gamma=0 (lower boundary, suggesting minimal pruning regularisation is preferred) and learning_rate=0.1 (upper boundary, suggesting a higher learning rate with more estimators might further improve performance). Results should therefore be interpreted with some caution as the true optimum may lie outside the tested ranges.

The selected bands within the 1140–1560 nm SWIR window likely relate to two distinct biochemical mechanisms. Bands around 1110–1130 nm reflect CH bond absorption in structural compounds such as cellulose and lignin [19], which break down under PVY-induced necrosis, while bands near 1450 nm capture OH bond absorption linked to leaf water content [19], consistent with the water stress caused by PLRV disrupting phloem transport [3, 7]. A per-class SHAP analysis was conducted to assess whether the model relies on different spectral bands for each disease (see Appendix G), confirming that PVY detection relies more on shorter SWIR wavelengths around 1033–1209 nm associated with structural tissue changes, while PLRV detection relies more on longer wavelengths around 1413–1589 nm associated with water content. That the model relies on bands from both regions suggests it captures both viral mechanisms simultaneously.

The two-stage feature engineering strategy, combining leaf-level ablation with plant-level aggregation, produced an

important insight about the role of spectral standard deviation. At the leaf level, spectral standard deviation seems to degrade classification because individual leaves naturally vary in angle, illumination, and developmental stage, making within-leaf variance a source of noise rather than signal [13]. At the plant level, however, the standard deviation across leaves becomes informative. It captures how much variation exists across the canopy, which seems to be a useful indicator of viral infection since infected plants tend to show more variation in their spectral responses across their leaves than healthy plants. This difference between noise and signal depending on the level of analysis has not been explicitly described in the existing literature.

The canopy size ablation showed another important finding: small leaves, which are often excluded from analyses under the assumption that they represent annotation artifacts or segmentation errors, seem to carry critical early-stage viral signatures. PLRV disrupts phloem transport and therefore particularly affects young developing tissue [3], which likely produces detectable spectral signatures in newly sprouted leaves that are absent or weaker in mature leaves. Including the full canopy therefore seems essential for early-stage detection. The upper-bound filters further revealed a growth-stage asymmetry: removing large leaves consistently hurt May performance but had little effect in June, where the most aggressive filter even slightly improved results. Since no single size filter improved both stages simultaneously, the no-filter baseline was retained.

The finding that the Top 10 and the Middle 10 LDA bands play different roles across growth stages has practical implications for sensor design. A fixed small subset of bands selected for one growth stage may not work well in another, which could explain inconsistencies reported in prior studies that used fixed narrow multispectral channels [4]. Furthermore, because the Top 20 LDA bands span a broader window from approximately 1140 to 1560 nm within the SWIR region, which is sensitive to leaf water content [7], any sensor designed for this task would need to cover this wider spectral range. The full Top 20 therefore seems necessary to span both growth stages and the full extent of discriminative wavelengths reliably.

Temporal Concept Drift is a significant practical limitation for automated disease detection in agriculture that has received limited attention in the existing HSI literature [22]. Because the model implicitly relies on physical size features, it seems to lose accuracy as plants grow and change between growth stages, regardless of how well the spectral features perform. For any real-world deployment this means that either separate models must be trained for each growth stage, or the training data must cover the full range of growth stages from the start [22].

The deployment evaluation (see Appendix E) revealed a further practical limitation: replacing ground truth masks with Mask R-CNN predicted masks caused a substantial drop in classification performance, from 88.5% to 61.5% in June and from 76.2% to 57.1% in May. While the predicted leaf count per plant closely matched the ground truth (56.0 versus 57.2 on average), detection quality remained imperfect. Roughly 30% of predicted leaves did not correspond to a real leaf and roughly 32% of real leaves were missed entirely, which likely introduced noise into the plant profile. These results suggest that the classification pipeline itself is not the main challenge for deployment, but rather that the segmentation quality is.

Finally, the LOVO results show the upper limit of what a single model trained across all varieties can achieve. The pipeline seems to identify discriminative spectral bands consistently across varieties, but the spectral baseline differences between varieties are large enough to confuse the classifier when no variety-specific examples are available in training. This is consistent with the findings of Griffel *et al.* [11].

5.2 Conclusion

The central question of this study was whether hyperspectral imaging combined with machine learning can be used to detect and classify PVY and PLRV at early infection stages across different potato varieties. The results show that this is partially possible. Within known varieties, the pipeline achieved strong classification performance, but generalisation to completely unseen varieties remains difficult, indicating that a single universal model is not yet sufficient for reliable cross-variety deployment.

Regarding spectral band selection, supervised Linear Discriminant Analysis (LDA)-guided band selection outperformed both the full spectrum and unsupervised alternatives, confirming that targeted dimensionality reduction is essential for classification performance. The discriminative bands span a broader spectral window than visual inspection alone would suggest.

For plant-level feature representation, aggregating leaf-level features into a plant profile improved classification by capturing canopy-level variation. The full canopy must be retained, as small leaves carry early-stage viral signatures that are lost when size filtering is applied.

Temporal generalisation proved to be a further challenge: models did not generalise well across growth stages, suggesting that growth stage-specific training is likely required for practical deployment. Segmentation quality was identified as a further obstacle for real-world use.

Together, these findings suggest that hyperspectral imaging with targeted band selection and plant-level aggregation is a promising direction for early virus detection, but that cross-variety generalisation and robust leaf segmentation remain the key challenges to address before field deployment becomes feasible.

5.3 Future Work

The findings of this study suggest several directions for future work. The spectral resolution ablation result suggests a direct hardware application: a lightweight multispectral camera calibrated to the 20 identified bands could eliminate the cost and complexity of a full hyperspectral sensor while retaining the most informative wavelengths. Since the Top 20 LDA-selected bands span a broader window from approximately 1140 to 1560 nm, such a sensor would need to cover this extended short-wave infrared range rather than

targeting only Range A and Range B. This would likely be more practical for large-scale field deployment than a line-scan camera. The performance drop observed when applying a model across growth stages could be addressed by either training a single model on data from both timepoints with growth stage included as an explicit input feature, or by implementing a calendar-based system that automatically selects the appropriate growth-stage-specific model. Improving Mask R-CNN segmentation quality, particularly reducing false positive leaf detections and missed leaves through better post-processing or additional training data, would be a useful step towards closing the gap between the controlled evaluation results and real-world deployment performance. Expanding the annotation dataset to cover more variety-disease combinations, particularly for varieties that currently only contain a single disease class, would likely increase the statistical power of LOVO evaluation. Using tools like Segment Anything to assist with annotation could also reduce the time needed to label new datasets. Testing whether the LDA-selected bands correspond to established vegetation indices such as the Water Band Index would help validate the biochemical interpretation. Finally, exploring training strategies where the model can be quickly updated with a small number of samples from a new variety, rather than requiring a full retraining from scratch, would directly address the cross-variety generalisation problem identified by the LOVO evaluation.

6 Ethical Impact Statement

This research uses hyperspectral images of potato plants collected under institutional data management guidelines. No human subjects or personal data were involved. All preprocessing steps, model configurations, and evaluation protocols are described in enough detail to allow full replication. Automated disease detection has clear benefits for society. Early identification of viral infections makes it possible to remove infected plants before the virus spreads, which can reduce crop losses and decrease the need for pesticide use. However, the risk of overconfidence in automated predictions should be acknowledged. The Temporal Concept Drift finding is particularly relevant here: a model trained on early growth stage data will likely give unreliable predictions when applied to late growth stage crops, which could cause disease to spread through missed detections. This pipeline is therefore intended to support, not replace, expert judgement in the field. It should be used as a tool that flags suspicious plants for further human inspection. Any deployment in practice should also involve close collaboration with end users to make sure the limitations of the model are clearly understood.

Acknowledgements

I would like to thank my supervisor Willem Dijkstra for his guidance and feedback throughout this internship. I would also like to thank the NHL Stenden Professorship in Computer Vision & Artificial Intelligence for providing access to the dataset, annotation tools, and computing resources that made this research possible. This project is financially supported by the European Union within the JTF-call project DigiAgro3.



References

[1] S. Savary et al., "The global burden of pathogens and pests on major food crops," *Nature Ecology & Evolution*, vol. 3, no. 3, pp. 430–439, 2019.

[2] B. Dupuis et al., "Economic impact of Potato Virus Y (PVY) in Europe," *Potato Research*, pp. 1–15, 2023.

[3] J. J. Couture, A. Singh, A. O. Charkowski, R. L. Groves, S. M. Gray, P. C. Bethke, and P. A. Townsend, "Integrating spectroscopy with potato disease management," *Plant Disease*, vol. 102, no. 11, pp. 2233–2240, 2018.

[4] T. Jia et al., "Detection of diseased potato plants with UAV hyperspectral imagery," in *Proc. 14th Workshop on Hyperspectral Imaging and Signal Processing (WHISPERS)*, IEEE, 2024, pp. 1–5.

[5] A. K. Mahlein et al., "Hyperspectral sensors and imaging technologies in phytopathology: State of the art," *Annual Review of Phytopathology*, vol. 56, pp. 535–558, 2018.

[6] M. R. Slaton, E. R. Hunt, and W. K. Smith, "Estimating near-infrared leaf reflectance from leaf structural characteristics," *American Journal of Botany*, vol. 88, no. 2, pp. 278–284, 2001.

[7] S. Junttila, T. Hölttä, N. Saarinen, V. Kankare, T. Yrttimaa, J. Hyyppä, and M. Vastaranta, "Close-range hyperspectral spectroscopy reveals leaf water content dynamics," *Remote Sensing of Environment*, vol. 277, p. 113071, 2022.

[8] J. Lapajne et al., "Comparison of selected dimensionality reduction methods for detection of root-knot nematode infestations in potato tubers using hyperspectral imaging," *Sensors*, vol. 22, no. 1, p. 367, 2022.

[9] G. Polder, P. M. Blok, H. A. C. de Villiers, J. M. van der Wolf, and J. J. Kamp, "Potato virus Y detection in seed potatoes using deep learning on hyperspectral images," *Frontiers in Plant Science*, vol. 10, p. 209, 2019.

[10] K. M. Gold et al., "Hyperspectral measurements enable pre-symptomatic detection and differentiation of contrasting physiological effects of late blight and early blight in potato," *Remote Sensing*, vol. 12, no. 2, p. 286, 2020.

[11] L. M. Griffel, D. Delparte, J. Whitworth, P. Bodily, and D. Hartley, "Evaluation of artificial neural network performance for classification of potato plants infected with potato virus Y using spectral data on multiple varieties and genotypes," *Smart Agricultural Technology*, vol. 3, p. 100101, 2023.

[12] K. Nagasubramanian et al., "Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems," *Plant Methods*, vol. 14, p. 86, 2018.

[13] S. Mazdeyasna, M. S. Arefin, A. Fales, S. J. Leavesley, T. J. Pfefer, and Q. Wang, "Evaluating normalization methods for robust spectral performance assessments of hyperspectral imaging cameras," *Biosensors*, vol. 15, no. 1, p. 20, 2025.

[14] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[15] P. Luo, Y. Yang, H. Zhang, M. Yi, X. Zhou, Y. Yang, C. Chen, M. Yan, and C. Jiang, "Identification of tobacco leaf diseases using hyperspectral imaging and machine learning with SHAP interpretability analysis," *Frontiers in Plant Science*, vol. 16, p. 1711972, 2025.

[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785–794.

[17] J. George, J. Yadav, and A. M. Nair, "Integrating LightGBM and XGBoost for robust plant disease classification: A homogenous stacking approach," in *Proc. International Conference on Advances in Data-Driven Computing and Intelligent Systems*, Springer Nature, 2024, pp. 409–422.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 2961–2969.

[19] P. K. Jayapal, R. Joshi, R. Sathasivam, B. V. Nguyen, M. A. Faqeerzada, S. U. Park, D. Sandanam, and B.-K. Cho, "Non-destructive measurement of total phenolic compounds in *Arabidopsis* under various stress conditions," *Frontiers in Plant Science*, vol. 13, p. 982247, 2022.

[20] B. C. Gao, "NDWI — A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sensing of Environment*, vol. 58, no. 3, pp. 257–266, 1996.

[21] E. R. Hunt and B. N. Rock, "Detection of changes in leaf water content using near- and middle-infrared reflectances," *Remote Sensing of Environment*, vol. 30, no. 1, pp. 43–54, 1989.

[22] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, 2014.

A Preprocessing Ablation

To verify that the combination of ℓ_2-normalisation and Savitzky-Golay derivative filtering is the most effective preprocessing strategy, four variants were compared at the leaf level using the full cropped spectrum and the plant-wise test set across both growth stages combined.

Table 10. Preprocessing ablation results at the leaf level (all dates combined, N = 47 test plants).

Variant	Acc	Prec	Rec	F1
A. Raw spectrum	61.7%	0.577	0.555	0.563
B. ℓ_2 norm only	61.7%	0.583	0.604	0.587
C. Derivative only	61.7%	0.558	0.555	0.555
D. ℓ_2 norm + derivative	**61.7%**	**0.607**	**0.601**	**0.603**

All four variants achieved identical accuracy (61.7%), indicating that preprocessing does not affect which plants are classified correctly at this level. However, the combined pipeline (D) achieved the highest macro F1 (0.603), driven by more balanced precision and recall across all three classes. Notably, derivative-only preprocessing (C) performed worse than ℓ_2 norm alone (B), suggesting that normalisation contributes more to class separation than the derivative transformation when applied in isolation. The combined pipeline was therefore retained as the default preprocessing strategy.

B Class Weighting Ablation

To assess the impact of balanced class weighting on baseline model performance, all three classifiers were evaluated under both weighted and unweighted conditions on the leaf-level baseline (full spectrum, no plant profile aggregation, no LDA selection). Tables 11 and 12 present the results for both evaluation strategies.

Table 11. LOVO cross-variety results with and without class weighting.

Weighting	Stage	Model	Acc	Prec	Rec	F1
Weighted	May	KNN	48.2%	0.403	0.354	0.344
Weighted	May	Rand. Forest	49.0%	0.397	0.348	0.325
Weighted	May	XGBoost	45.6%	0.432	0.363	0.360
Unweighted	May	KNN	48.2%	0.403	0.354	0.344
Unweighted	May	Rand. Forest	48.7%	0.462	0.357	0.343
Unweighted	May	XGBoost	47.8%	0.439	0.357	0.345
Weighted	June	KNN	40.5%	0.396	0.352	0.342
Weighted	June	Rand. Forest	41.1%	0.434	0.347	0.317
Weighted	June	XGBoost	42.7%	0.422	0.357	0.337
Unweighted	June	KNN	40.5%	0.396	0.352	0.342
Unweighted	June	Rand. Forest	41.2%	0.432	0.350	0.324
Unweighted	June	XGBoost	42.9%	0.437	0.362	0.338

Table 12. Plant-wise results with and without class weighting.

Weighting	Stage	Model	Acc	Prec	Rec	F1
Weighted	May	KNN	51.2%	0.475	0.403	0.415
Weighted	May	Rand. Forest	57.6%	0.562	0.454	0.468
Weighted	May	XGBoost	52.8%	0.517	0.480	0.486
Unweighted	May	KNN	51.2%	0.475	0.403	0.415
Unweighted	May	Rand. Forest	55.5%	0.536	0.444	0.458
Unweighted	May	XGBoost	53.9%	0.531	0.447	0.461
Weighted	June	KNN	52.2%	0.525	0.513	0.515
Weighted	June	Rand. Forest	59.5%	0.597	0.579	0.579
Weighted	June	XGBoost	62.4%	0.618	0.609	0.609
Unweighted	June	KNN	52.2%	0.525	0.513	0.515
Unweighted	June	Rand. Forest	59.2%	0.592	0.576	0.575
Unweighted	June	XGBoost	61.5%	0.616	0.596	0.597

The differences between weighted and unweighted configurations are small across all models and both growth stages, indicating that class weighting has limited impact at the leaf level under the full-spectrum baseline. The relative ranking of classifiers is preserved in both conditions, confirming that XGBoost's selection as the primary classifier is not dependent on the weighting choice. Balanced class weighting was retained for all main experiments as it produced consistently equal or better F1 scores, particularly for the minority class.

C PVY Penalty Ablation

To justify the selection of the 1.5× PVY penalty multiplier, a systematic grid search was conducted over seven penalty values ranging from 1.0× to 3.0×, using the Top 20 LDA bands and optimised XGBoost hyperparameters on the plant-wise test set. Table 13 present the results for both growth stages. The 1.5× multiplier achieved the best combined performance across both growth stages, matching the highest F1 in June (0.866) while maintaining the same May performance as the unpenalised 1.0× baseline (F1: 0.792). Notably, at 1.0× the June PVY recall dropped to 0.50, meaning half of all PVY-infected plants were missed, while at 1.5× PVY recall in June improved to 0.88 with no loss in overall accuracy. Penalties above 1.5× consistently degraded May performance without further improving June results. The 1.5× multiplier was therefore selected as the optimal balance between PVY recall and overall classification performance.

Table 13. PVY penalty ablation results for both Early (May) and Late (June) Growth Stages.

	Early Growth Stage (May)				Late Growth Stage (June)			
Penalty	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
1.0×	**76.2%**	**0.790**	**0.800**	**0.792**	80.8%	0.830	0.770	0.763
1.25×	71.4%	0.700	0.740	0.717	88.5%	0.870	0.850	0.849
1.5×	**76.2%**	**0.790**	**0.800**	**0.792**	**88.5%**	**0.870**	**0.870**	**0.866**
1.75×	66.7%	0.670	0.710	0.683	80.8%	0.760	0.760	0.758
2.0×	71.4%	0.720	0.770	0.732	84.6%	0.820	0.800	0.804
2.5×	66.7%	0.690	0.680	0.673	84.6%	0.820	0.800	0.804
3.0×	61.9%	0.630	0.600	0.602	88.5%	0.910	0.840	0.854

D Per-Variety Mean Derivative Spectra

Figures 4 and 5 show the mean first-derivative normalised reflectance spectra per variety for the early growth stage (21-05-2025) and late growth stage (13-06-2025) respectively.

Each subplot shows the per-class mean spectrum for Healthy, PVY, and PLRV where available, with Range A (1080–1150 nm) and Range B (1330–1400 nm) shaded. Dashed vertical lines indicate the valley position within each range per class. The consistent x-axis shifts in Range A across most varieties motivated its selection as a key discriminative region, while Range B shows limited visual class separation across varieties.

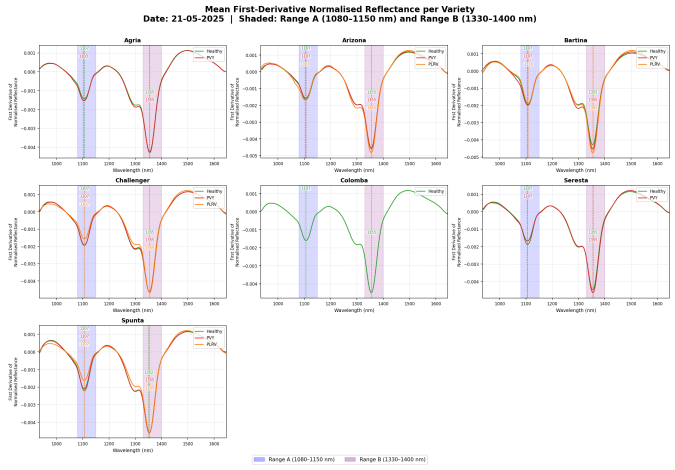


Fig. 4. Mean first-derivative normalised reflectance per variety, early growth stage (21-05-2025).

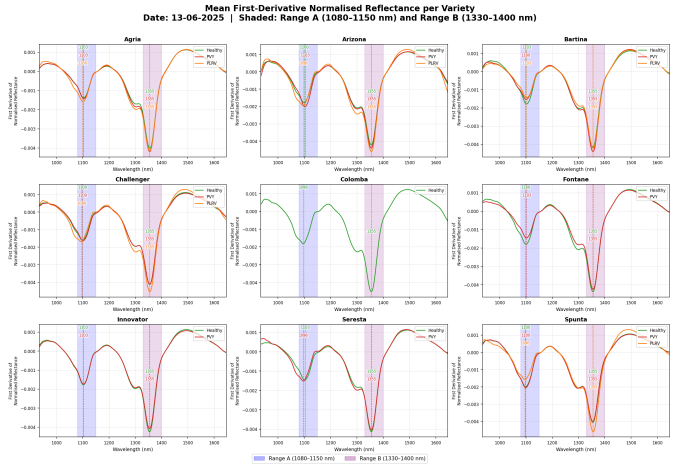


Fig. 5. Mean first-derivative normalised reflectance per variety, late growth stage (13-06-2025).

E Mask R-CNN Deployment Evaluation

To evaluate the pipeline under realistic deployment conditions, the Mask R-CNN was used to generate predicted masks for the test set, which were then used as input to the XGBoost classifier instead of the ground truth annotations. Table 14 shows the performance comparison between both conditions.

Table 14. Classification performance using ground truth masks versus Mask R-CNN predicted masks on the test set.

Growth Stage	Mask Source	Acc	Prec	Rec	F1
Early (May)	Ground truth annotations	76.2%	0.792	0.792	0.792
Early (May)	Mask R-CNN	57.1%	0.420	0.470	0.429
Late (June)	Ground truth annotations	88.5%	0.866	0.866	0.866
Late (June)	Mask R-CNN	61.5%	0.570	0.560	0.562

Performance dropped when using predicted masks in both growth stages. In May, accuracy fell from 76.2% to 57.1% and F1 from 0.792 to 0.429, with the drop in recall particularly notable since PVY (class 2) was never correctly predicted, likely because only 6 PVY plants were present in the May test set, making it highly sensitive to any misclassification. In June, accuracy fell from 88.5% to 61.5% and F1 from 0.866 to 0.562, a smaller but still substantial drop. Segmentation evaluation revealed that while the predicted leaf count per plant closely matched the ground truth (56.0 versus 57.2 on average), detection quality remained imperfect. Mask precision was 0.691 and recall was 0.676, meaning roughly 30% of predicted leaves did not correspond to a real leaf and roughly 32% of real leaves were missed entirely. This mismatch introduces noise into the plant profile and accounts for the performance drop. This result shows that the classification pipeline itself is not the challenge for deployment, the segmentation quality is.

F Per-Variety LOVO Results

To assess whether cross-variety generalisation varies across potato varieties, the final LOVO evaluation results are broken down per variety for both growth stages in Tables 15 and 16. Colomba is excluded from both tables as it contains only a single disease class and therefore does not qualify for LOVO evaluation.

Table 15. Per-variety LOVO results, early growth stage (May), sorted by F1. Accuracy, precision, recall, and F1 are computed per variety; the Mean ± SD row reflects variation across the 6 qualifying test varieties.

Variety	Train	Test	Acc	Prec	Rec	F1
Challenger	51	12	58.3%	0.639	0.556	0.584
Spunta	51	12	66.7%	0.533	0.556	0.517
Bartina	51	12	66.7%	0.476	0.611	0.503
Seresta	57	6	66.7%	0.556	0.444	0.489
Arizona	51	12	33.3%	0.400	0.278	0.306
Agria	57	6	33.3%	0.167	0.222	0.190
Mean ± SD			**54.2 ± 15.1%**	**0.462 ± 0.151**	**0.445 ± 0.147**	**0.431 ± 0.137**

Table 16. Per-variety LOVO results, late growth stage (June), sorted by F1. Accuracy, precision, recall, and F1 are computed per variety; the Mean ± SD row reflects variation across the 8 qualifying test varieties.

Variety	Train	Test	Acc	Prec	Rec	F1
Spunta	65	12	91.7%	0.952	0.889	0.908
Challenger	65	12	83.3%	0.867	0.889	0.850
Agria	66	11	63.6%	0.422	0.600	0.492
Arizona	65	12	58.3%	0.405	0.500	0.447
Fontane	71	6	50.0%	0.389	0.333	0.356
Seresta	71	6	50.0%	0.250	0.500	0.333
Innovator	71	6	33.3%	0.200	0.333	0.250
Bartina	68	9	22.2%	0.222	0.222	0.222
Mean ± SD			**56.5 ± 21.8%**	**0.463 ± 0.271**	**0.533 ± 0.234**	**0.482 ± 0.244**

The results reveal substantial variation across varieties, as reflected by the high standard deviations (15.1% in May, 21.8% in June). In June, Spunta and Challenger generalised very well (F1 0.908 and 0.850 respectively), suggesting their

spectral disease signatures are similar enough to other varieties in the training set to transfer. At the other extreme, Bartina (F1 0.222) and Innovator (F1 0.250) generalised poorly, indicating that their spectral baselines differ enough from other varieties to confuse the classifier. In May, performance varied without a clear pattern across the six qualifying varieties.

G Per-Class SHAP Analysis

To assess whether the model relies on different spectral features for PVY and PLRV, a per-class SHAP analysis was conducted using TreeExplainer on XGBoost models trained on all 140 plant profiles per growth stage. These models are not cross-validated and are fitted on the full available data, so the SHAP values are intended for model explanation only and should not be interpreted as classification performance estimates.

Figure 6 shows the mean absolute SHAP value per feature and per class for both growth stages. Features are ordered by wavelength from left to right; the dashed vertical line separates spectral features from morphological features. Figures 7 and 8 show the corresponding per-class beeswarm plots for the early and late growth stages. Each dot represents one plant; its position on the x-axis indicates the direction and magnitude of the SHAP contribution toward predicting that class, and its colour indicates the raw feature value (red = high, blue = low).

In May, the three classes rely on partially distinct spectral features. For the Healthy class, the dominant features are morphological: Perimeter mean (mean $|\text{SHAP}| = 0.436$) and Total pixels (0.424), with high perimeter and total pixel count pushing predictions toward Healthy (beeswarm directions: $+0.73$ and $+0.58$ for high feature values). The spectral feature Mean 1244 nm also ranks highly for Healthy (0.366), where high reflectance in that region pushes predictions away from Healthy (-0.25 for high values), consistent with healthy tissue showing lower water-stress-related absorption in the SWIR.

For PVY, the top features are spectral: Std 1413 nm (0.343), Std 1205 nm (0.285), Mean 1033 nm (0.261), and Mean 1205 nm (0.218). The beeswarm shows that high Std 1413 nm strongly pushes predictions away from PVY (-0.63 for high values), suggesting that plants with high canopy-level variance in the 1413 nm region are less likely to be PVY-infected. These wavelengths, particularly around 1110–1130 nm, correspond to CH bond absorption in structural compounds such as cellulose and lignin [19], which is consistent with PVY-induced cell damage.

For PLRV, the dominant features are Total pixels (0.756) and Mean 1557 nm (0.727). The beeswarm shows a clear directional pattern: high Total pixels pushes strongly away from PLRV (-0.81), while high Mean 1557 nm pushes toward PLRV ($+0.77$). This suggests PLRV-infected plants tend to have fewer total leaf pixels combined with elevated reflectance at 1557 nm. This falls within the OH bond absorption region linked to leaf water content [19], consistent with PLRV disrupting phloem water transport [3]. Leaf count (0.192) also ranks in the PLRV top 5, with high leaf count pushing away from PLRV (-0.29), which is plausible since PLRV stunts plant development and reduces canopy size.

In June, the most important features change substantially. For the Healthy class, Perimeter mean becomes the dominant feature (0.858), with high perimeter strongly predicting Healthy ($+0.98$ for high values). This morphological dominance suggests that by June, leaf shape is a more reliable indicator of healthy growth than spectral reflectance, possibly because both viruses cause visible leaf deformation by this stage. Solidity mean (0.266) and Leaf count (0.417) are the next most important features.

For PVY, Perimeter mean (0.463) and Solidity mean (0.321) are the top features, but with inverted directions relative to Healthy: both high Perimeter mean and Solidity mean push away from PVY (-0.58 and -0.61). Std 1219 nm (0.208) and Eccentricity std (0.193) also contribute, with high Std 1219 nm pushing toward PVY ($+0.20$).

For PLRV in June, spectral features become dominant again, particularly Mean 1589 nm (0.492), Mean 1420 nm (0.483), Mean 1560 nm (0.350), and Mean 1413 nm (0.281). The beeswarm shows that high reflectance at 1589 nm pushes toward PLRV ($+0.33$), while high Mean 1420 nm pushes strongly away (-0.58). These wavelengths overlap with known OH bond absorption regions, consistent with the 1450 nm absorption band reported in the literature [19] and are consistent with the water stress signature of PLRV [3, 7]. The concentration of PLRV-relevant features in the SWIR water absorption window, combined with the morphological features dominating Healthy and PVY classification, suggests the model captures both structural and water-related disease signals in the late growth stage.

Taken together, the SHAP analysis confirms that PVY and PLRV are not detected through identical spectral signatures, at least in the early growth stage. PVY detection relies more on shorter SWIR wavelengths around 1033–1209 nm associated with structural tissue changes, while PLRV detection relies more on longer SWIR wavelengths around 1413–1589 nm associated with water content. This partial distinction aligns with the different biological mechanisms of the two viruses and supports the interpretation that the 1140–1560 nm window selected by LDA captures both structural and water-related disease signals simultaneously. In the late growth stage, PVY and PLRV converge more on morphological features, with PLRV retaining a distinct spectral signature in the water absorption region.

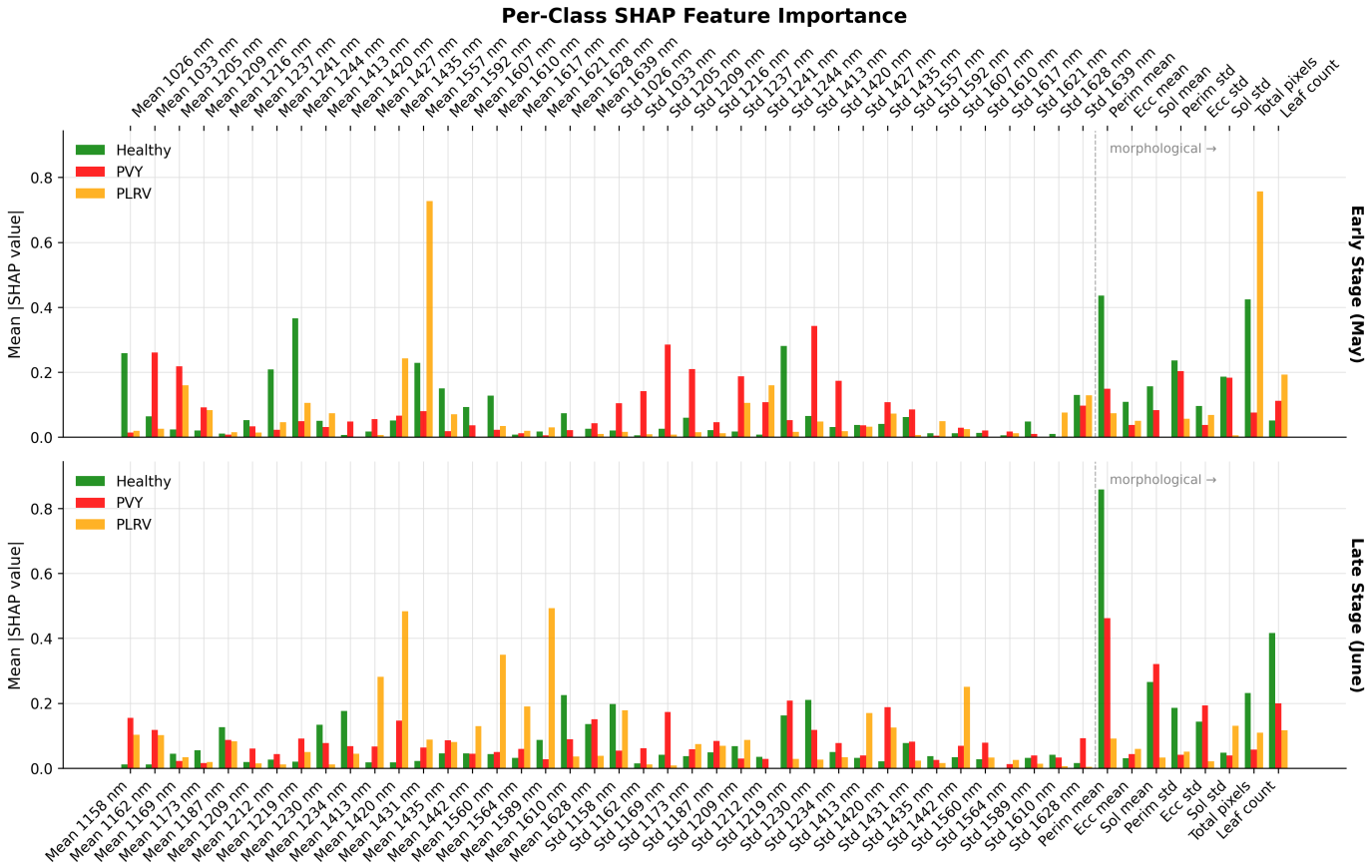


Fig. 6. Per-class mean absolute SHAP feature importance for the early growth stage (May, top) and late growth stage (June, bottom). Green = Healthy, red = PVY, orange = PLRV. Features are ordered by wavelength; the dashed vertical line separates spectral from morphological features.

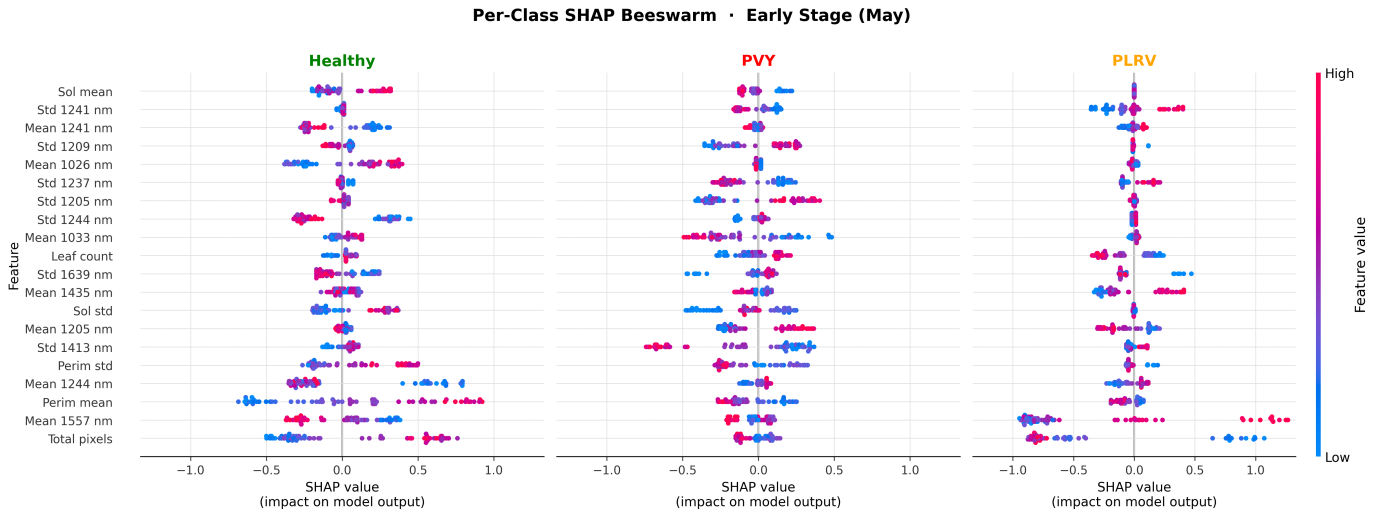


Fig. 7. Per-class SHAP beeswarm plots for the early growth stage (May). Each panel shows the top 20 features ranked by mean absolute SHAP value for one class. Each dot represents one plant; position on the x-axis indicates the direction and magnitude of the SHAP contribution (right = pushes model toward predicting this class; left = pushes away), and colour indicates the raw feature value (red = high, blue = low).

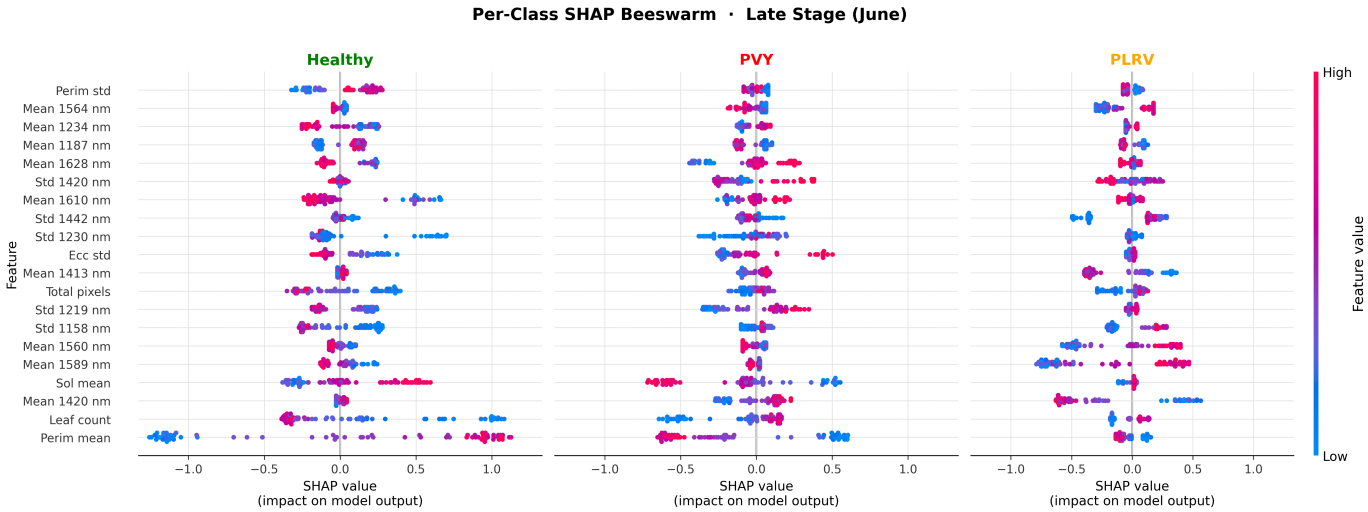


Fig. 8. Per-class SHAP beeswarm plots for the late growth stage (June). Each panel shows the top 20 features ranked by mean absolute SHAP value for one class. Each dot represents one plant; position on the x-axis indicates the direction and magnitude of the SHAP contribution (right = pushes model toward predicting this class; left = pushes away), and colour indicates the raw feature value (red = high, blue = low).