Received: 26 December 2023 | Accepted: 29 June 2024

DOI: 10.1111/ppa.13973

Plant Pathology An International Journal edited by the British Society for Plant Pathology WILEY

ORIGINAL ARTICLE

# Aerial low-altitude remote sensing and deep learning for in-field disease incidence scoring of virus yellows in sugar beet

Nathan Okole[1] | Facundo R. Ispizua Yamati[1] | Roxana Hossain[1] | Mark Varrelmann[1] | Anne-Katrin Mahlein[1] | Rene H. J. Heim[1,2]

[1]Institute of Sugar Beet Research (IfZ), Göttingen, Germany

[2]Photogrammetry & Robotics Lab, Institute for Photogrammetry, University of Bonn, Bonn, Germany

**Correspondence**
Nathan Okole, Institute of Sugar Beet Research (IfZ), Holtenser Landstrasse 77, DE 37079 Göttingen, Germany.
Email: okole@ifz-goettingen.de

**Funding information**
Deutsche Forschungsgemeinschaft, Grant/Award Number: EXC 2070-390732324

## Abstract

This study investigates the potential of high-resolution (<0.5 cm/pixel) aerial imagery and convolutional neural networks (CNNs) for disease incidence scoring in sugar beet, focusing on two important aphid-transmitted viruses, beet mild yellowing virus (BMYV) and beet chlorosis virus (BChV). The development of tolerant sugar beet cultivars is imperative in the context of increased disease management concerns due to the ban on neonicotinoids in the European Union. However, traditional methods of disease phenotyping, which rely on visual assessment by human experts, are both time-consuming and subjective. Therefore, this study assessed whether aerial multispectral and RGB images could be harnessed to perform automated disease ratings comparable to those performed by trained experts. To this end, two variety trials were conducted in 2021 and 2022. The 2021 dataset was used to train and validate a CNN model on five cultivars, while the 2022 dataset was used to test the model on two cultivars different from those used in 2021. Additionally, this study tests the use of transformed features instead of raw spectral bands to improve the generalization of CNN models. The results showed that the best CNN model was the one trained for BMYV on RGB images using transformed features instead of conventional raw bands. This model achieved a root mean square error score of 11.45% between the model and expert scores. These results indicate that while high-resolution aerial imagery and CNNs hold great promise, a complete replacement of human expertise is not yet possible. This research contributes to an innovative approach to disease phenotyping, driving advances in sustainable agriculture and crop breeding.

KEYWORDS
convolutional neural networks, phenotyping, unmanned aerial vehicle, virus yellows

## 1 | INTRODUCTION

In recent years, the ban on neonicotinoid insecticides within the European Union has resulted in the resurgence of aphid-transmitted plant viral diseases that were previously well controlled in sugar beet (Hossain et al., 2021). Virus yellows (VY) is a disease complex caused by different virus species that are transmitted by the green peach aphid (*Myzus persicae*). These virus species include two poleroviruses from the *Solemoviridae* family, beet mild yellowing virus (BMYV) and beet chlorosis virus (BChV), a closterovirus from the *Closteroviridae* family, beet yellows virus (BYV) and a potyvirus from the *Potyviridae* family, beet mosaic virus (BtMV) (Hossain et al., 2021). Of these, BMYV and BChV

  

have emerged as significant threats to sugar beet cultivation in Europe, causing yield losses of up to 23% and 24%, respectively (Hossain et al., 2021). In the absence of neonicotinoid insecticides for seed coating and subsequent disease control, the development of resistant cultivars is the most sustainable solution (Hossain et al., 2021). However, developing tolerant sugar beet cultivars is hindered by the phenotyping process, which has become a major bottleneck (Mahlein et al., 2019).

Phenotyping plays a pivotal role in the development of tolerant and resistant cultivars, as it helps in identifying plants with desirable traits (Yang et al., 2017). Traditionally, this process has been time-consuming and resource-intensive as it relies on tedious visual examination carried out by trained experts (Mahlein et al., 2019). Despite providing insightful data, this method is inherently subjective and can be inconsistent across experts (Bock et al., 2020). Therefore, the term 'ground reference' is preferred to 'ground truth' when referring to ground data collected by trained experts, as the latter implies that ground data are error-free (Justice & Townshend, 1981). The urgent need to accelerate and enhance the objectivity of breeding resistant sugar beet cultivars requires a paradigm shift in phenotyping approaches (Mahlein et al., 2019).

Contributing to such a shift might be the rapid evolution of unmanned aerial vehicle (UAV) technologies, integrated with increasingly sophisticated data analysis approaches including machine learning (ML) and deep learning (DL), which provides an innovative solution to the challenges posed by traditional phenotyping methods (Soori et al., 2023). UAVs equipped with high-resolution cameras (<0.5 cm resolution) have the potential to capture detailed information on plant health status swiftly and objectively. By leveraging ML and DL techniques, these images can be systematically and automatically analysed, eliminating the subjectivity associated with human evaluations (Bock et al., 2022; Mahlein et al., 2019).

Several studies have explored the potential of UAVs for stress phenotyping applications. For example, a study by Trapp (2015) compared tolerant and susceptible recombinant inbred lines of dry bean under terminal drought conditions using high-resolution multispectral images. Plot-to-plot comparison of the green normalized difference vegetation index (GNDVI) with yield data resulted in a strong correlation ($r = 0.79$, $p = 0.01$). Another study by Chivasa et al. (2020) evaluated 25 maize varieties grown in a trial inoculated with maize streak virus. Multispectral images were analysed using random forest (RF) models. Correlations between the UAV-derived data and manual maize streak virus scores were significant ($r = 0.74–0.84$). Barreto et al. (2023), with overall accuracy values of up to 85.8%, established a pipeline based on ML methods to extract disease-relevant parameters for Cercospora leaf blight in sugar beet phenotyping. Ispizua Yamati et al. (2024) also used UAV multispectral and RGB images combined with different ML and DL approaches, including convolutional neural network (CNN) models to score Rhizoctonia crown and root rot severity on sugar beet. In this study, precision values ranging from 0.73 to 0.85 could be achieved for Rhizoctonia crown and root rot scoring.

While the promise of UAVs and ML/DL for disease phenotyping is compelling, it is essential to acknowledge that their effectiveness varies depending on the particular model and pathosystem considered (Dhaka et al., 2021). As the ML and DL approaches are empirical modelling procedures, large volumes of data are needed to capture often complex, dynamic and variable phenomena. Furthermore, while DL techniques such as CNNs are considered cutting-edge image analysis methods, cases exist where they have been outperformed by older and less complex ML methods (Li et al., 2020). Therefore, each plant–pathogen interaction may necessitate tailored approaches for accurate disease assessment.

This study aims to develop a CNN-based approach for automatic incidence scoring of virus yellows disease in sugar beet and to evaluate the effectiveness of this proposed approach. In this study, we explore the utilization of transformed features, as opposed to raw spectral bands, to enhance the generalization of CNNs. To the best of our knowledge, this approach has not been extensively investigated in the context of CNNs. Furthermore, unlike most previous CNN studies, the model developed in this research was rigorously assessed on a completely independent dataset, reinforcing the credibility of our accuracy assessments. This study ultimately contributes to the improvement of phenotyping approaches for the development of more tolerant sugar beet cultivars in the post-neonicotinoid era.

## 2 | MATERIALS AND METHODS

### 2.1 | Study area and experimental design

Data from two different trials were harnessed for the analysis presented in this study. In 2021, a virus yellows (VY) variety trial was conducted on five sugar beet (*Beta vulgaris*) genotypes (Figure 1). The experimental site was located in Sieboldshausen, Germany (51°28′13″ N, 09°54′20″ E). The trial was arranged in a two-factorial design with four replicates. The two factors were the inoculation variant comprising three inoculation strategies (not inoculated, inoculated with BMYV and inoculated with BChV) and the genotype (from susceptible to tolerant) comprising five genotypes from three different suppliers (Table 1). Seeds were sown on 3 April 2021, in plots (8 × 1.5 m) containing three rows of sugar beets, resulting in approximately 100 plants in each plot. To prevent unintentional virus spread by insects, border rows and control plots were sprayed with insecticides. To produce viruliferous aphids (*M. persicae*) for the inoculation, healthy aphids were placed on virus-infected mother plants that were produced in sufficient quantities in advance in the greenhouse. The aphids were placed on plants infected with the individual viruses for at least 48 h to acquire the virus. The inoculation of plants was conducted on 31 May 2021, with 10 viruliferous wingless aphids being transferred to 3% of the plants per plot. Inoculated plants were at the phenological stage corresponding to plants with four unfolded and fully developed leaves (BBCH 14; Meier et al., 2009).

Data from another VY cultivar trial, conducted in 2022, on two sugar beet genotypes different from those used in 2021, was

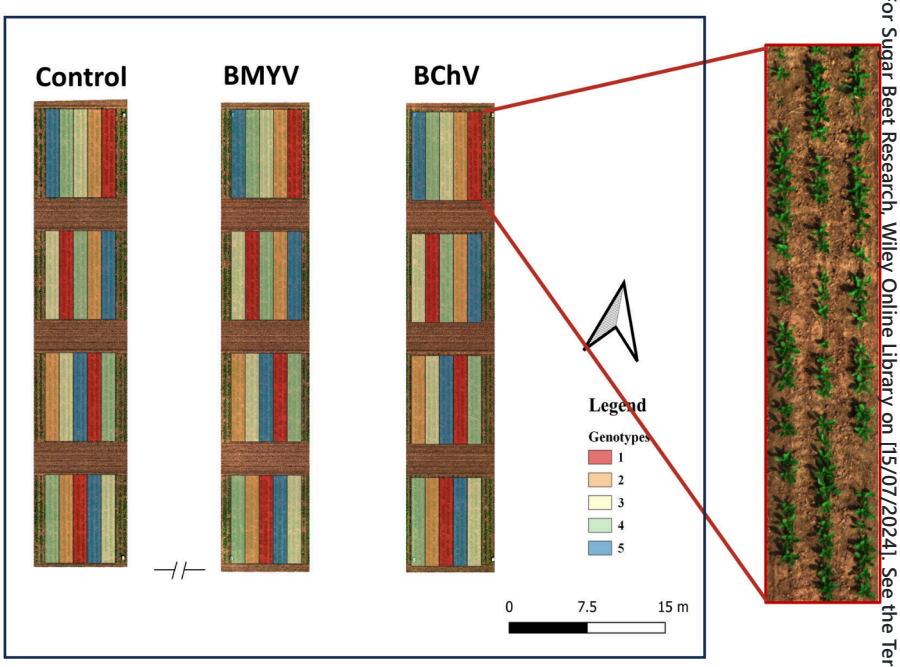


**FIGURE 1** Composed RGB image of the first experimental site on 8 June 2021 under sunny conditions. Different inoculation variants (Control, beet mild yellowing virus [BMYV], beet chlorosis virus [BChV]) as well as genotypes (coded from 1 to 5) are depicted. More information on the genotypes is given in Table 1. The discontinuity sign in the image indicates that the 'Control' treatment was further away from the two other treatments (about 75 m).

**TABLE 1** Information about the sugar beet genotypes used in this study.

| Genotype | Breeder's code | Provider | Putative reaction |
|---|---|---|---|
| 1 | BMYV2021 | KWS | Tolerant |
| 2 | ST-VY220 | Strube | Tolerant |
| 3 | TolSV2020 | SES | Tolerant |
| 4 | ST-VY120 | Strube | Susceptible |
| 5 | ZR 2313 | SES | Susceptible |

harnessed to validate the developed models. The trial included a tolerant (T) and a susceptible (S) genotype (Figure 2). The inclusion of both susceptible and tolerant genotypes provided a high-variability dataset to develop a more robust model (Tasdizen et al., 2018). The experiment was carried out at an independent site located in Harste, Germany (51°36′00″ N, 09°52′00″ E). For this second experiment, only data for BMYV was available. Again, the trial was arranged in a two-factorial design, but this time the factors were the time of inoculation and the genotypes, with four replicates for each combination. The time of inoculation included four levels: (a) not inoculated, (b) inoculated at the phenological stage BBCH 12 (mid-May), (c) inoculated at row closure (mid-June) and (d) inoculated 3 weeks after row closure (mid-July). An overview of the experimental design is presented in Table 2. For this experiment, plots were sown on 31 March 2022. The plots had the same dimensions as those in the 2021 experiment and were separated by a 6 m distance to avoid contaminations between different virus species tested on the same field location.

## 2.2 | Visual assessment of VY

In both years, the monitoring of virus symptoms was performed from June to September (Table 3). The visual assessment was carried out by an expert standing in front of the plots. During the initial observation, the number of plants exhibiting yellowing symptoms was estimated. This was followed by an accurate counting. Disease incidence (DI) was then determined using Equation 1.

$$DI = (\text{No. of plants with symptoms} / \text{Total no. of plants assessed}) \times 100$$

Both numbers (DI values from estimate and counting) were then compared (data not shown) and yielded nearly similar values, allowing for subsequent observations to rely solely on estimates of the number of plants displaying symptoms. The yellowing symptoms caused by BMYV and BChV could be distinguished from other types of yellowing, including those caused by insects and wilt, as they typically manifest initially at the leaf margin but subsequently spread over the entire leaf, as described by Hossain et al. (2021).

## 2.3 | Image acquisition

To allow for automatic scoring, aerial imaging was conducted in addition to manual scoring. The UAV platform was the Matrice 210 (SZ, DJI Technology Co.). Multispectral images with 1280×960 pixels were captured using the Altum multispectral snapshot camera (MicaSense). The Altum has five multispectral bands including blue (B: 475 nm centre wavelength, 32 nm bandwidth), green (G: 560 nm, 27 nm), red (R: 668 nm, 14 nm), red edge (RE: 717 nm, 12 nm) and near-infrared (NIR: 842 nm, 57 nm) and an integrated long-wave thermal infrared (TIR) sensor (band range 8–14 μm). Images were captured simultaneously on the six different cameras in 16-bit raw GeoTIFF format. The capture was done at an altitude of 15 m above the soil level, with 70% forward and lateral overlap and an average flight speed of 0.4 m/s, resulting in an average ground sampling distance (GSD) of 4 mm. A grey reference

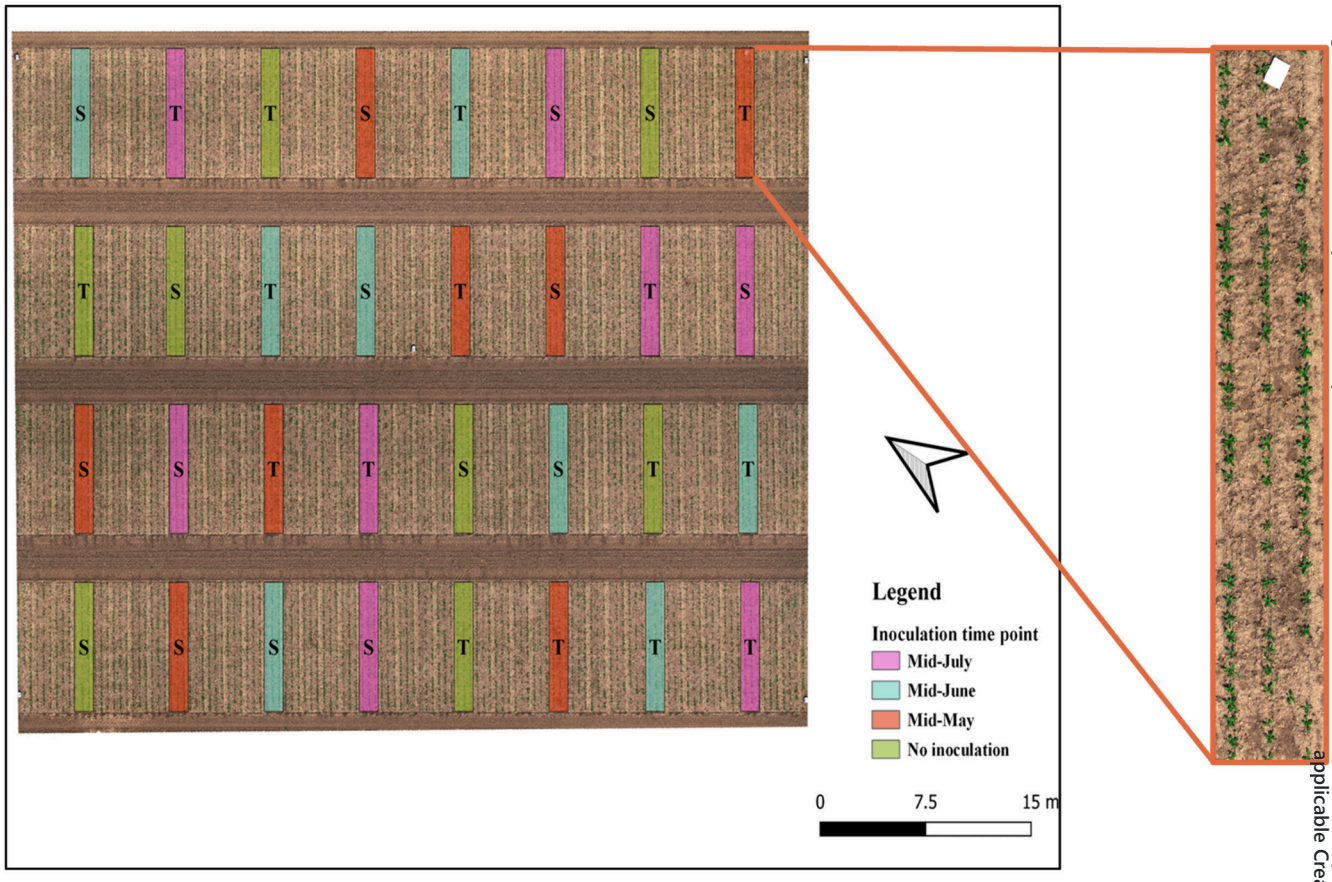


FIGURE 2 Composed RGB image of the second experimental site on 20 May 2022 under sunny conditions. Different inoculation time points as well as genotypes are depicted. 'T' refers to the putative tolerant sugar beet genotype, while 'S' refers to the putative susceptible one. More information on the time point labels is found in Table 2.

TABLE 2 Summary of treatments for the 2022 trial.

| Genotype | Inoculation stage | Inoculation date |
|---|---|---|
| S | No inoculation | – |
| | BBCH 12 | 16 May 2022 (46 das) |
| | Row closure | 22 Jun 2022 (83 das) |
| | 3 weeks after row closure | 14 Jul 2022 (107 das) |
| T | No inoculation | – |
| | BBCH 12 | 16 May 2022 (46 das) |
| | Row closure | 22 Jun 2022 (83 das) |
| | 3 weeks after row closure | 14 Jul 2022 (107 das) |

Abbreviations: das, days after sowing; S, susceptible; T, tolerant.

panel, provided by the camera manufacturer, was imaged directly before and after each flight for radiometric calibration using the empirical line method (Aasen et al., 2018). To ensure precise georeferencing and co-registration, ground control points were installed at the corner points of the field and measured with a real-time kinematics (RTK) positioning system for referencing the generated map according to its actual geographic location. Eventually, georeferencing errors of less than 0.1 cm were achieved for all flights. A total of six flights in 2021 and five in 2022 were conducted as reported in Table 3. Flight missions were always within ±8 days of visual scoring dates to allow for comparison. All flight missions took place close to solar noon between 11:00 AM and 1:00 PM, with sunny weather and rare occurrence of clouds.

## 2.4 | Image processing

### 2.4.1 | Orthomosaic generation

The UAV imagery was processed using Agisoft Metashape v. 1.8.3 (Agisoft LLC). The workflow for orthomosaic generation involved importing images, aligning them, generating a dense point cloud, adding ground control points, performing radiometric correction, building a digital elevation model (DEM), generating an orthomosaic and exporting it as a tagged image file. For all these steps, preconfigured settings from the Metashape Professional User Manual v. 1.8 (Agisoft LLC, 2022) were used with slight modifications as reported in Table S1.

### 2.4.2 | Extraction of images of single plants

Before training a supervised CNN classification model to distinguish between healthy and diseased plants, individual sugar beet



13653059, 2024, 0, Downloaded from https://bsppjournals.onlinelibrary.wiley.com/doi/10.1111/ppa.13973

**TABLE 3** Visual scoring and unmanned aerial vehicle (UAV) mission dates.

| Year | Visual scoring | UAV mission | Days of difference | Crop phenological stage |
|---|---|---|---|---|
| 2021 | – | 8 Jun 2021 (10 dai) | – | BBCH 37 |
| | 28 Jun 2021 (30 dai) | 28 Jun 2021 (30 dai) | 0 | BBCH 39 |
| | 13 Jul 2021 (45 dai) | 13 Jul 2021 (45 dai) | 0 | BBCH 39 |
| | 28 Jul 2021 (60 dai) | 3 Aug 2021 (66 dai) | +6 | BBCH 39 |
| | 9 Sep 2021 (103 dai) | 2 Sep 2021 (96 dai) | −7 | BBCH 39 |
| | – | 29 Sep 2021 (123 dai) | – | BBCH 39 |
| 2022 | – | 20 May 2022 (50 das) | – | BBCH 17 |
| | 22 Jun 2022 (83 das) | 23 Jun 2022 (84 das) | +1 | BBCH 39 |
| | 7 Jul 2022 (98 das) | – | – | BBCH 39 |
| | 22 Jul 2022 (113 das) | 19 Jul 2022 (110 das) | −3 | BBCH 39 |
| | 4 Aug 2022 (126 das) | – | – | BBCH 39 |
| | 17 Aug 2022 (139 das) | 24 Aug 2022 (146 das) | +7 | BBCH 39 |
| | 1 Sep 2022 (154 das) | – | – | BBCH 39 |
| | 15 Sep 2022 (168 das) | 23 Sep 2022 (176 das) | +8 | BBCH 39 |

*Note*: Phenological stage BBCH 17 refers to the stage where plants have seven unfolded leaves, BBCH 37 refers to the stage where leaves cover 70% of the ground and BBCH 39 refers to the stage where leaves cover 90%–100% of the ground.

Abbreviations: dai, days after inoculation; das, days after sowing.

plants had to be detected and then cropped from the orthomosaic. Our approach, developed in Python v. 3.8 (Python Software Foundation, 2021), used a modified version of the workflow by Günder et al. (2022) for individual plant detection. Modification included the removal of the Hough transform-based row detection algorithm, which in our case resulted in the omission of many sugar beet plants. After plant centroid detection using the approach by Ispizua Yamati et al. (2024), individual sugar beet plants were cropped around the centroid position. A ground distance of 22.5 cm was considered on either side of the plant to crop a 45 cm square image around the centroid of the plant. This image size allows the incorporation of an adult plant's largest approximated surface area (Ispizua Yamati et al., 2024). For further processing, these images were assigned to classes according to the variety, the inoculation variant and the ID of the plot they were cropped from.

### 2.4.3 | CNN modelling and training

A CNN was used for the classification. The architecture of the CNN model was similar to that of AlexNet (Krizhevsky et al., 2017), which proved to be suitable for other disease classification tasks (Saleem et al., 2019), with slight modifications as presented in Table S2. To build the training dataset, two replicate plots out of four were considered for each genotype. The other two plots were left for validation. Images that were cropped from inoculated plots were assessed visually for symptoms and were labelled as 'Diseased' if yellowing and necrotic symptoms were present. The same was done for non-inoculated plots and the images were assigned to the 'Healthy' class. Eventually, the dataset was made of 5801 images, of which 3352 belonged to the class 'Healthy', 1722 to the class 'BChV' and 727 to the class 'BMYV'. To improve the generalization ability and to balance the model, data augmentation was applied (Wong et al., 2016). This included random rotations, shifts and slight zooms and eventually 5000 images were obtained for each class. In total, four different models were trained, including two for each disease (i.e., BChV and BMYV).

For each disease, one model was constructed using only the RGB bands of the Altum camera and the other model was built by considering all bands of the Altum camera, except the thermal. These models are later referred to as RGB-based and multispectral-based models, respectively. At this point, a choice of the type of training features had to be made. After a first evaluation using raw spectral bands (i.e., red, green, blue, …), it was noticed that the models did not achieve a good validation accuracy (results not shown). Therefore, the final models were trained on preprocessed features instead of raw spectral bands. The latter approach showed a more robust validation accuracy. For RGB-based models, three features were computed, namely hue (H; Equation 2), saturation (S; Equation 3) and green leaf index (GLI; Equation 4). For multispectral-based models, four features were computed, among which were the hue and saturation as before, and the optimized soil-adjusted vegetation index (OSAVI; Equation 5) and the normalized difference red-edge (NDRE) index (Equation 6).

$$H=\begin{cases}0, & \text{if} \max(R,G,B)-\min(R,G,B)=0\\ 60\times\dfrac{G-B}{\max(R,G,B)-\min(R,G,B)}, & \text{if} \max(R,G,B)=R\\ 60\times\left(\dfrac{B-R}{\max(R,G,B)-\min(R,G,B)}+2\right), & \text{if} \max(R,G,B)=G\\ 60\times\left(\dfrac{R-G}{\max(R,G,B)-\min(R,G,B)}+4\right), & \text{if} \max(R,G,B)=B\end{cases} \quad (2)$$

$$S = \begin{cases} 0, & \text{if } \max(R,G,B) = 0 \\ \dfrac{\min(R,G,B)}{\max(R,G,B)}, & \text{otherwise} \end{cases} \quad (3)$$

$$GLI = \frac{2G - R - B}{2G + R + B} \quad (4)$$

$$OSAVI = \frac{NIR - R}{NIR + R + 0.16} \quad (5)$$

$$NDRE = \frac{NIR - RE}{NIR + RE} \quad (6)$$

The training process was carried out using the adaptive moment estimation (Adam) optimizer, with an initial learning rate (Lr), a batch size and a maximum number of epochs of 0.001, 256 and 100, respectively. Throughout the training, two metrics were continuously monitored: validation loss, which gauges how well the model is performing and how close its predictions are to the actual values, and validation accuracy, which indicates the proportion of correct predictions relative to the total predictions made on the validation dataset. The damping factor for the Lr was set to 0.1 with a patience of three epochs, so that the Lr was lowered by factor 0.1 after three consecutive epochs with no improvement in the validation loss. Early stopping was also applied with a patience of 7 to avoid overfitting.

The architecture and weights of the proposed CNN model were finally stored as a hierarchical data format (H5) file for further validation processes.

#### 2.4.4 | Visual assessment of CNN models

While deep learning models are often regarded as black box models (Bilbrey et al., 2020), it is possible to reveal and visualize the features in the image that have relevance for the classification. For a visual explanation of models in this study, the gradient-weighted class activation mapping (Grad-CAM) method was used (Selvaraju et al., 2017). Here, the final convolutional layer is converted into a localization map that indicates significant image areas for predicting a specific class (in this situation, 'diseased'). These localization maps were then plotted as heat maps to visually assess whether the model selects areas that can be biologically linked to the examined disease symptoms.

#### 2.4.5 | Accuracy assessment (validation and testing)

As the data was pooled and analysed from two different datasets, the accuracy assessment of the model was also done based on this background. A thorough accuracy assessment should include validation and testing. During the validation phase, the model's performance is evaluated on a separate dataset not used during training, ensuring that it generalizes well to unseen data. This step is vital to detect overfitting, where the model performs exceptionally well on the training data but poorly on new data (Vabalas et al., 2019). The testing phase is the final step in assessing accuracy. In this phase, the model is put to the ultimate test by evaluating its performance on an entirely independent dataset, distinct from both the training and validation datasets. This ensures that the model's accuracy remains robust and trustworthy, offering a true reflection of its real-world capabilities (Malebary & Khan, 2021).

First, validation was done on the two plots per treatment that were left untouched during model training in 2021 for each assessment date. Secondly, testing was performed using the dataset collected on different genotypes in a different location and setting in 2022. The CNN model's prediction (i.e., healthy or diseased) was recorded for all plant images from the same plot, and disease incidence (%) in each plot was calculated as the ratio of images classified as containing diseased plants. The resulting scores for each plot were then stored in a data frame, along with the ground reference score for the same plot, for further agreement analyses between the two scores.

### 2.5 | Statistical analysis

An analysis of variance (ANOVA) was carried out to test the effect of the genotype on disease incidence across different blocking factors (i.e., inoculation variant, inoculation time point). The agreement between the model-derived score and the ground reference score was assessed by fitting a linear function, using metrics such as the slope, the Pearson correlation coefficient ($r$), the coefficient of determination ($R^2$) and the root mean squared error (RMSE) as indicators of the performance. In addition, Lin's concordance correlation coefficient (CCC) was computed to account for both precision and bias (Lin, 1989). All statistical analyses were performed in R software (R Core Team, 2016), and all analyses involving comparisons of means were carried out using the 'agricolae' package (De Mendiburu, 2014). Additionally, disease development was analysed by fitting disease progress models to the data using the 'epifitter' package in R (Alves & Del Ponte, 2021). Four models of disease progression were fitted to the incidence data. These models were the Gompertz, logistic, exponential and monomolecular models. The coefficient of determination was used to select the best-fitting model (in our case, the logistic and Gompertz models for the years 2021 and 2022, respectively). The infection rate (IR) value derived from the model was then used to compare different disease evolution dynamics.

## 3 | RESULTS

### 3.1 | Disease development in different trials

First, only the expert-based incidence and disease progression are outlined, without linking them to the CNN-based scoring. The results of the expert-based disease incidence scoring for the years

2021 and 2022 are presented in Figures 3 and 4, respectively. During the rating period in year 2021, the different genotypes showed a heterogenous response to BMYV and BChV. Although the most tolerant genotype to BMYV was also the most tolerant to BChV, the two diseases showed slightly different dynamics over the growing season. For BMYV, differences in expert-based scores were observed as early as 30 days after infection (dai) ($p = 0.046$), while for BChV, the first significant differences across genotypes were only detected at 45 dai ($p = 0.050$). Later in the season (60 dai), no more significant differences could be observed across genotypes for BMYV ($p = 0.272$), whereas for BChV, the difference was still significant ($p = 0.039$). Towards the end of the growing season, disease incidence was severe for all genotypes except for BMYV on genotype 1, which had a significantly lower incidence compared to other genotypes ($p = 0.045$).

For the year 2022 (Figure 4), only BMYV was used to inoculate one susceptible and one tolerant genotype at different time points, which were approximately 1 month apart from each other. As expected based on infection rate (IR) values (Table 4), the disease developed faster for the treatments that were inoculated earlier (mid-May). For this inoculation time point, the disease developed faster in the susceptible genotype (IR = 0.112) compared to the tolerant one (IR = 0.036). Although at the end of the season the disease incidence was the same for both the tolerant and the susceptible genotypes reaching 100% ($p = 0.185$), the progression of the disease was different. Delaying the inoculation time point by approximately 1 month (mid-June) had a significant effect on the disease dynamics. The infection rate was dramatically reduced for both the susceptible (IR = 0.023) and the tolerant (IR = 0.015) genotype compared to early inoculation (mid-May). A significant difference between the two genotypes could only be seen towards the end of the season ($p = 0.027$, $p < 0.001$ and $p = 0.027$ at 126, 154 and 168 days after sowing [das], respectively). Further delaying the inoculation time point by 3 weeks (mid-July) made the difference between the

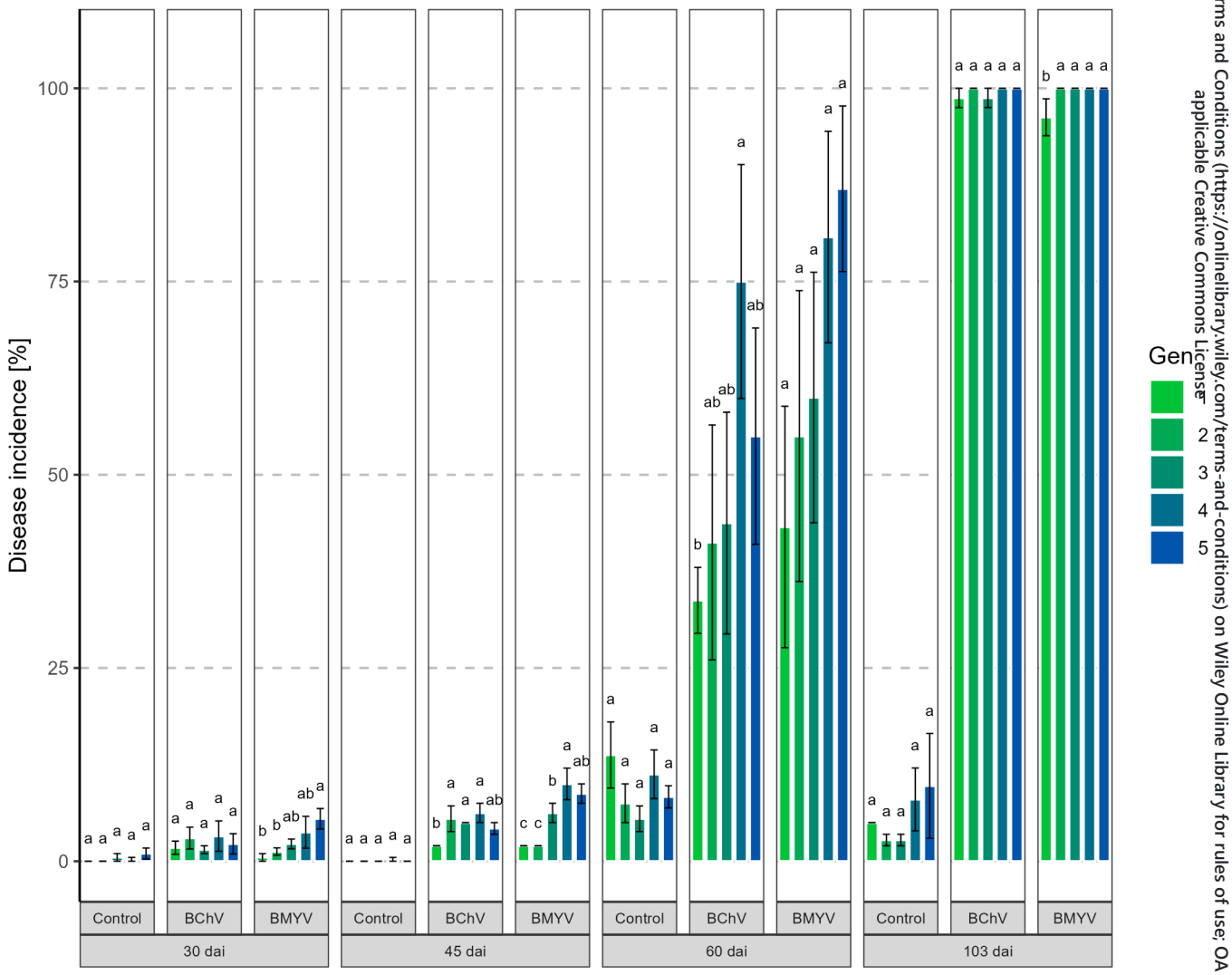


FIGURE 3 Temporal development of disease on different genotypes following beet chlorosis virus (BChV) and beet mild yellowing virus (BMYV) inoculation (experiment in 2021). The numbers in the legend and the corresponding colour relate to the different genotypes described in Table 1. Dai, days after inoculation; Gen, genotype.

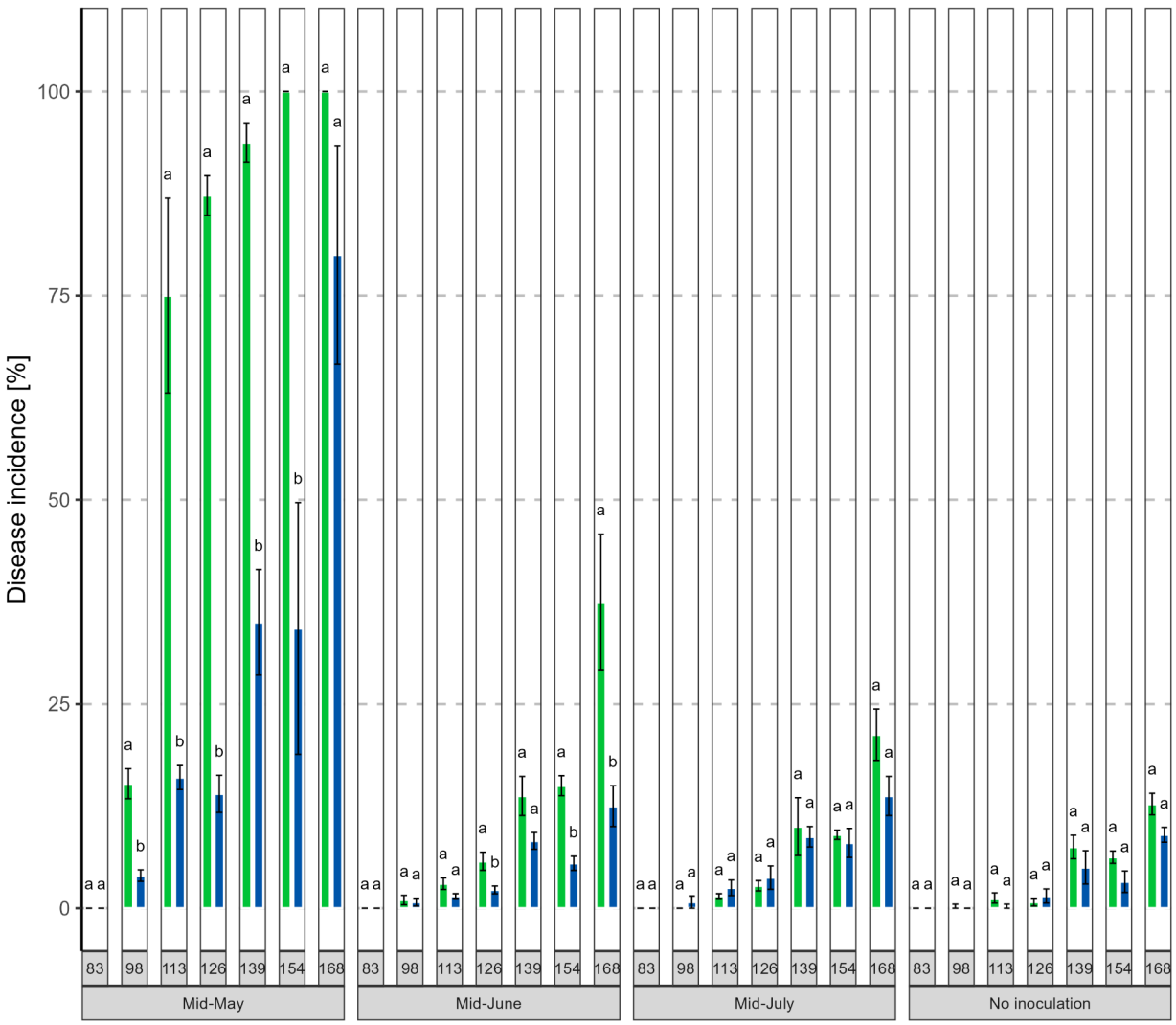


FIGURE 4 Temporal development of beet mild yellowing virus (BMYV) in 2022 in a susceptible (S) and tolerant (T) sugar beet variety inoculated at different time points ('No inoculation', 'Mid-May', 'Mid-June' and 'Mid-July', as explained in Table 2). The numbers on the *x*-axis indicate the number of days after sowing. Gen, genotype.

tolerant and the susceptible genotype unnoticeable ($p > 0.05$ for all dates). Furthermore, the progression of the disease score for this last inoculation time point was almost similar to noninoculated treatments (IR = 0.016–0.022). This anticipated reduction in infection rates is likely to be due to the shorter time left for disease development and less conducive climatic conditions for disease transmission by the aphid when inoculation occurs late in the growing season. The variability in the dataset due to different infection rates ultimately permitted a comprehensive assessment of the CNN model.

## 3.2 | Evaluation of the predictive ability of the CNN models

In this section, differences between RGB and multispectral data will be reported. The accuracy and loss curves during training of the CNN models used to discriminate healthy plants from those infected with BChV and BMYV are reported in Figure 5. The models were built on RGB or multispectral images and achieved, on average, high overall accuracy (>95%) on the validation set and low loss (<0.2) after approximately 10 epochs, regardless of the disease. However, for BChV, the multispectral-based model showed better performance as indicated by its lower loss compared to the RGB-based model. Conversely, for BMYV, both the RGB-based and the multispectral-based model showed similar performance.

By analysing, the activation maps of the last convolutional layer produced by the Grad-CAM method (Figure 6), it was observed that the multispectral-based model for BChV relied on visible necrotic and chlorotic windows (highlighted by white circles in the image) to make its predictions. In contrast, the RGB-based CNN model relied on other features besides visible symptoms (highlighted by red circles). For the BMYV disease, the RGB-based model relied on visible symptoms to make its predictions, while the multispectral-based model sometimes based its decision on visible symptoms and other

**TABLE 4** Disease development models fitted to disease incidence data for different sugar beet genotypes in both trials across different blocking factors (i.e., inoculation variant, inoculation time point).

| Year | Inoculation variant | Genotype | Model | IR | IR_se | IR_ci_lwr [95%] | IR_ci_upr [95%] | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| 2021 | BChV | 1 | Logistic | 0.123 | 0.017 | 0.051 | 0.194 | 0.965 |
| | BChV | 2 | Logistic | 0.149 | 0.017 | 0.077 | 0.221 | 0.976 |
| | BChV | 3 | Logistic | 0.120 | 0.008 | 0.084 | 0.156 | 0.990 |
| | BChV | 4 | Logistic | 0.148 | 0.017 | 0.077 | 0.219 | 0.976 |
| | BChV | 5 | Logistic | 0.154 | 0.015 | 0.087 | 0.220 | 0.980 |
| | BMYV | 1 | Logistic | 0.119 | 0.019 | 0.039 | 0.199 | 0.953 |
| | BMYV | 2 | Logistic | 0.163 | 0.019 | 0.080 | 0.247 | 0.973 |
| | BMYV | 3 | Logistic | 0.152 | 0.011 | 0.102 | 0.201 | 0.989 |
| | BMYV | 4 | Logistic | 0.144 | 0.014 | 0.083 | 0.205 | 0.981 |
| | BMYV | 5 | Logistic | 0.141 | 0.020 | 0.053 | 0.229 | 0.960 |
| | Control | 1 | Logistic | 0.093 | 0.060 | −0.166 | 0.351 | 0.545 |
| | Control | 2 | Logistic | 0.084 | 0.055 | −0.152 | 0.319 | 0.539 |
| | Control | 3 | Logistic | 0.045 | 0.056 | −0.194 | 0.284 | 0.245 |
| | Control | 4 | Logistic | 0.053 | 0.031 | −0.082 | 0.187 | 0.586 |
| | Control | 5 | Logistic | 0.058 | 0.061 | −0.204 | 0.319 | 0.310 |
| **Year** | **Inoculation time point** | **Genotype** | **Model** | **IR** | **IR_se** | **IR_ci_lwr [95%]** | **IR_ci_upr [95%]** | **$R^2$** |
| 2022 | No inoculation | S | Gompertz | 0.017 | 0.002 | 0.011 | 0.022 | 0.921 |
| | No inoculation | T | Gompertz | 0.017 | 0.002 | 0.011 | 0.023 | 0.922 |
| | Mid-May | S | Gompertz | 0.112 | 0.011 | 0.083 | 0.141 | 0.952 |
| | Mid-May | T | Gompertz | 0.036 | 0.006 | 0.021 | 0.050 | 0.889 |
| | Mid-June | S | Gompertz | 0.023 | 0.002 | 0.017 | 0.029 | 0.955 |
| | Mid-June | T | Gompertz | 0.015 | 0.002 | 0.009 | 0.022 | 0.886 |
| | Mid-July | S | Gompertz | 0.022 | 0.003 | 0.015 | 0.029 | 0.935 |
| | Mid-July | T | Gompertz | 0.016 | 0.002 | 0.010 | 0.023 | 0.896 |

*Note*: Four models were fitted for each case, namely Gompertz, logistic, exponential and monomolecular, and only the best model was retained. The genotypes for 2021 are coded from 1 to 5 and are further described in Table 1.

Abbreviations: BMYV, beet mild yellowing virus; BChV, beet chlorosis virus; IR, infection rate; IR_ci_lwr (95%) and IR_ci_upr (95%), lower and upper limits of the 95% confidence interval for the IR estimate, respectively; IR_se, standard error of the infection rate; S, susceptible; T, tolerant.

times on irrelevant features in the image (i.e., shadows and healthy tissues).

### 3.3 | Comparison between automatic scoring of the disease using a CNN and expert-based scoring

Figure 7 outlines the comparison between the expert-based disease scores and the model-based disease scores for different diseases and different input image types derived from the year 2021 (validation dataset from the same experimental setup as training data). In general, all model-based scores showed a good agreement with expert-based scores (RMSE < 16%, CCC > 0.90), with the models for BChV being on average more accurate (RMSE < 11%, CCC > 0.94). As expected from the visual assessment of the CNN models for BChV, a slightly better performance was obtained from multispectral images (RMSE = 10.5%, CCC = 0.954) compared to RGB images (RMSE = 10.89%, CCC = 0.946). In contrast, the performance of the CNN models for BMYV did not follow the predictions of the visual assessment. In this case, multispectral images (RMSE = 15.35%, CCC = 0.904) showed a slightly higher performance than RGB images (RMSE = 15.7%, CCC = 0.901), although the activation map showed that the multispectral-based model made its decision based on biologically unexplainable features. All CNN models except the RGB-based model for BChV tended to underestimate disease incidence.

### 3.4 | Testing the CNN models

The CNN models for BMYV were further evaluated using a dataset collected in the year 2022, from a different location and with a different experimental setup. The results of this evaluation are presented in Figure 8. Although the multispectral-based model performed well in the first year (RMSE = 15.4%, CCC = 0.904), this situation was inverted in the second year (RMSE = 42.9%, CCC = 0.384) compared to the RGB-based model, which showed an even better



13653059, 2024, 0, Downloaded from https://bsppjournals.onlinelibrary.wiley.com/doi/10.1111/ppa.13973

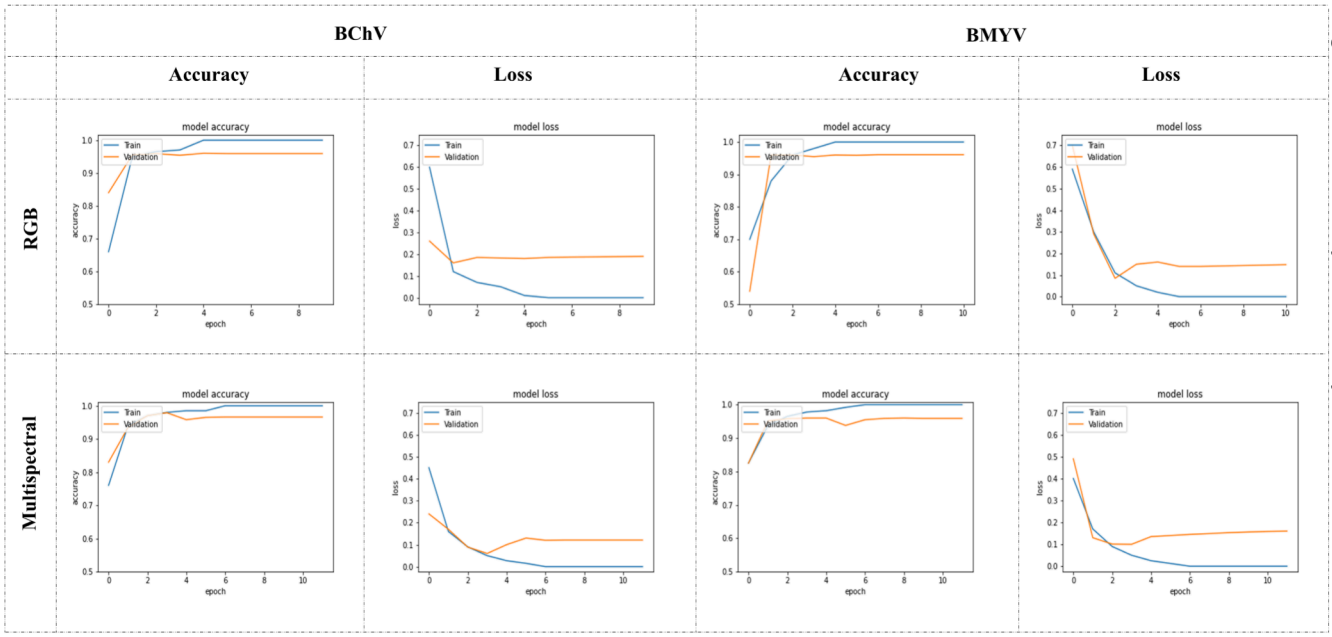


FIGURE 5 Accuracy and loss curves during training of CNN classification models for beet chlorosis virus (BChV) and beet mild yellowing virus (BMYV).

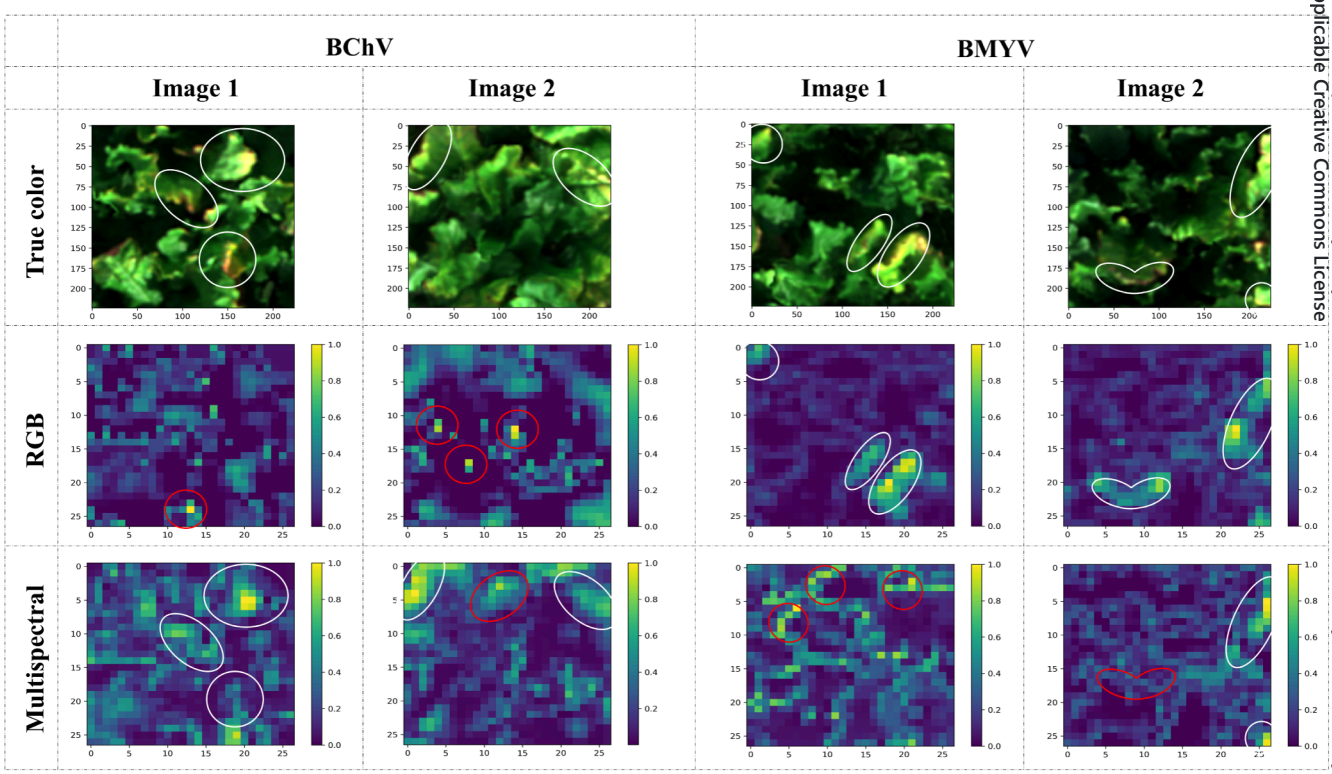


FIGURE 6 Activation map of the last convolutional layer of CNN classification models for beet chlorosis virus (BChV) and beet mild yellowing virus (BMYV). Two example images are shown for each disease. Regions where visual symptoms align with activation maps are circled in white, while areas displaying discrepancies, such as detection of shadow and healthy tissues as symptoms or failure to detect specific symptoms, are circled in red.

performance in the second year (RMSE = 15.7%, CCC = 0.901 in year 2021 vs. RMSE = 11.4%, CCC = 0.911 in year 2022). This observation aligns with the prediction of the activation maps, which showed that the RGB-based models learned relevant features for the disease. However, although RGB-based models showed high accuracy, there were still some underestimations (6/128) and overestimations (1/128)

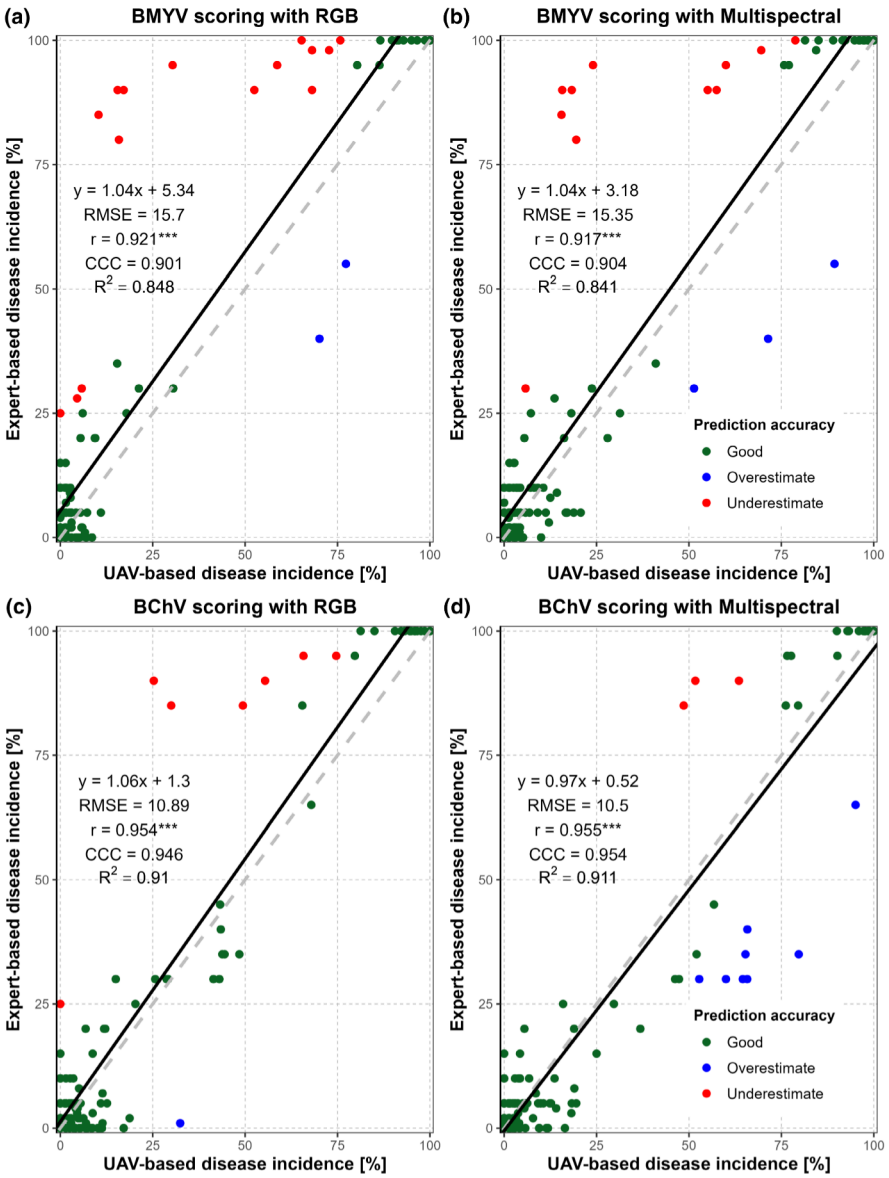


**FIGURE 7** Comparison between expert-based disease incidence and unmanned aerial vehicle (UAV)-based disease incidence in the year 2021 (first experiment). The coefficient of determination ($R^2$), the correlation coefficient ($r$), the root mean square error (RMSE) and Lin's concordance correlation coefficient (CCC) are indicated in the figure. Plots where the score was correctly estimated by the model (less than 15% difference) are shown in green, while overestimated and underestimated ones are shown in blue and red, respectively. The 1:1 line indicating perfect agreement is shown with a grey dashed line. (a, b) Beet mild yellowing virus (BMYV) with RGB and multispectral images, respectively. (c, d) Beet chlorosis virus (BChV) with RGB and multispectral images, respectively.

## 4 | DISCUSSION

This research was undertaken to assess the potential of UAVs and CNNs for disease incidence scoring in sugar beet phenotyping, specifically for BMYV and BChV. A dataset collected in 2021 was used to train a CNN model, while another dataset collected in 2022, in a different experimental setup and with only BMYV, was used to test the trained CNN model in real-world conditions. In addition, the use of RGB images versus multispectral images as input data to the CNN model, as well as the use of transformed features for training CNN models, were compared.

The research outcomes shed light on the potential of UAVs for advancing disease phenotyping precision in sugar beet breeding. Our most effective CNN model, trained to score disease incidence on RGB images for BMYV in 2021, showed good generalization abilities when evaluated on a hold-out validation dataset in the same year (RMSE = 15.7%, CCC = 0.904). In addition to the conventional practice of testing the model on a hold-out dataset, our model was tested on a distinct dataset from 2022, characterized by a different experimental setup and with different genotypes and exhibited even better performance (RMSE = 11.43%, CCC = 0.911).

Achieving high performance on previously unseen data is a noteworthy outcome. Such an outcome can be attributed to several factors, including (a) an initial training dataset with substantial variability (Therrien & Doyle, 2018), (b) the application of image processing techniques, such as data augmentation (Thanapol et al., 2020), which artificially augments the training dataset's variability, or (c) the training of the model on transformed features (Liu et al., 2020) that are less dependent on conditions that are specific to the training dataset. In our case, the model's robustness results primarily from the utilization of data augmentation techniques, such as random image rotation, flipping, zoom in and zoom out, and, to a greater extent, from the utilization of transformed features instead of the original spectral bands. When the model was exclusively trained on the original spectral bands, it demonstrated strong performance on the

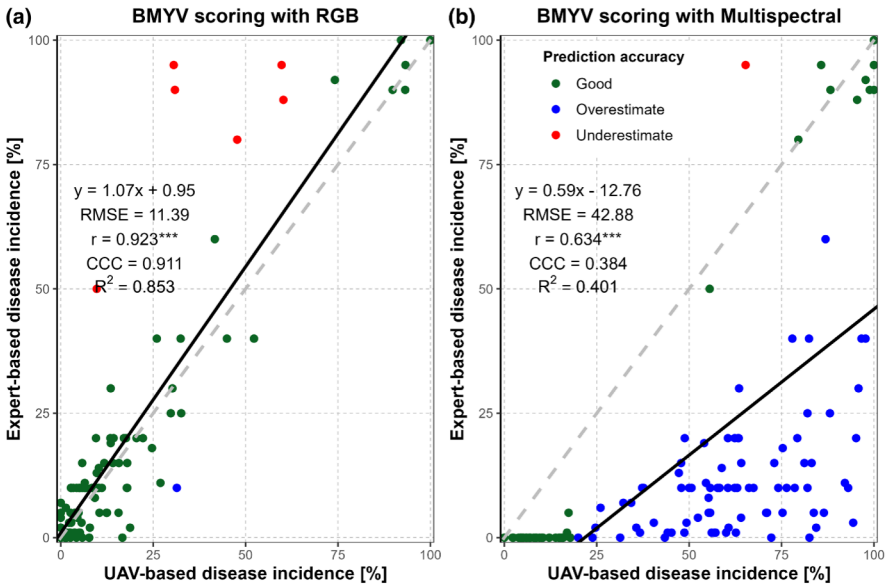


**FIGURE 8** Comparison between expert-based disease incidence and unmanned aerial vehicle (UAV)-based disease incidence in the year 2022 (second experiment in a different location). The coefficient of determination ($R^2$), the correlation coefficient ($r$), the root mean square error (RMSE) and Lin's concordance correlation coefficient (CCC) are indicated in the figure. Plots where the score has been correctly estimated by the model (less than 15% difference) are shown in green, while overestimated and underestimated ones are shown in blue and red, respectively. The 1:1 line indicating perfect agreement is shown with a grey dashed line. (a, b) Beet mild yellowing virus (BMYV) with RGB and multispectral images, respectively.

hold-out subset from the same year. However, it was unable to generate accurate predictions and generalize effectively to the dataset from the following year (results not shown). In contrast, employing transformed features such as the HSV colour space (hue and saturation only) and vegetation indices, which exhibit minimal sensitivity to variations in weather and lighting conditions (Purkis & Klemas, 2011), yielded a more resilient model capable of robust generalization to unseen data. Although feature transformation has been extensively explored and appreciated in the field of conventional machine learning, it remains relatively underexploited in the field of deep learning (Gowda & Yuan, 2019). A study by Krishnaswamy Rangarajan and Purushothaman (2020) showed that using the RGB colour space was better than using the YCbCr colour space to classify eggplant diseases, with overall accuracies (OAs) of 94.7% and 87.8%, respectively, when using pictures taken in laboratory conditions where illumination effects are minimal. When images taken in field conditions were used instead, both models showed similar performance (OA = 99.4%).

The ability to achieve good results on previously unseen data, particularly in the context of changing experimental conditions and genotypic diversity, reflects the promise of data augmentation techniques and adequate use of transformed features when training a CNN model (Gowda & Yuan, 2019). While data augmentation is well known and commonly applied, feature transformation has some potential to further improve the accuracy of CNN models, specifically in the fields of disease detection and diagnosis. In this study, only colour space transformation and vegetation indices were explored as feature transformation approaches. Other feature transformation approaches such as texture features are also worth investigating to build more robust models. Texture features contain valuable information related to the spatial arrangement of pixels in an image (Muneer & Fati, 2020). For disease detection, texture features can be highly informative, especially in cases where textural variations are key indicators of disease presence.

In addition, this study emphasizes that although CNNs are often regarded as black boxes (i.e., they produce outstanding results with often high accuracy, but the process leading to the result cannot be logically/mathematically tracked; Schramowski et al., 2020), they should not be treated as such (Ras et al., 2022). Neglecting the need to comprehend how CNNs achieve their predictions before deploying them in practical applications can lead to inaccuracies in the outputs. In fact, it is highly probable that a model making accurate predictions based on incorrect assumptions will fail to generalize effectively to unseen data (Ispizua Yamati et al., 2024).

In this study, we evaluated the generalization ability of the different CNN models using the Grad-CAM method. This method uses the final convolutional layer to generate a rough localization map, which is subsequently evaluated visually. While quantitatively evaluating the models' performance (measured by RMSE and CCC) for BMYV detection, it appeared that the model trained on multispectral data (Figure 7a, RMSE = 15.4, CCC = 0.904) performed slightly better than its RGB-trained counterpart (Figure 7b, RMSE = 15.7, CCC = 0.901). However, during the visual evaluation using the Grad-CAM method (Figure 6), it became evident that the RGB-based model's predictions exhibited greater robustness. This observation was corroborated when both models were tested on previously unseen data. In

agreement with the results of the visual assessment, the RGB-based model (Figure 8a, RMSE = 11.45, CCC = 0.911) performed significantly better than its multispectral-based counterpart (Figure 8b, RMSE = 42.9, CCC = 0.384).

This result underscores that relying solely on metrics such as RMSE or CCC for comparing or selecting the best CNN model can be misleading. It can sometimes happen that the model with the better RMSE or CCC is inferior when performing visual assessments and therefore poorly generalizes to unseen data as observed in the present study. The same observation was made by Lin et al. (2022), who used the Grad-CAM method to visualize the results of three different CNN models trained to detect grapevine foliar diseases. Their results showed that using the F1-score metric to select the best model would have resulted in selecting the wrong model in terms of the visualization map. In fact, the model with the highest F1-score in their study had activation maps that were not focused on disease symptoms. Surprisingly, the model with the lowest F1-score had its activation maps precisely focused on disease symptoms, which ultimately resulted in the highest overall accuracy score on the test set. It is therefore imperative to stop treating CNNs like black boxes and to perform visual assessment before their practical utilization.

This study also highlights the potential of multispectral and RGB imaging for disease phenotyping. Multispectral imagery captures a wider range of spectral information compared to standard RGB imaging, enabling the detection of subtle disease-related changes that may go unnoticed with the naked eye or traditional cameras. This was particularly evident in the context of BChV, where the symptoms were not as prominent as in BMYV (Figure 3). Conversely, in the context of BMYV, where infected plants showed more conspicuous yellowing symptoms, the benefit of using multispectral images instead of RGB images could not be established. This finding demonstrates that the decision whether to use multispectral or RGB imagery should depend on the epidemiology of the specific disease as well.

BMYV and BChV, while both affecting sugar beet, exhibit unique characteristics in terms of disease progression and symptom manifestation. The research emphasizes that a one-size-fits-all approach is not adequate when it comes to UAVs and CNNs for disease incidence scoring. Instead, it highlights the necessity of customizing methodologies, sensor configurations and data processing techniques to suit the distinct behaviours of different pathogens.

Finally, the research findings demonstrate the potential advantages of using UAVs in disease phenotyping and scoring, particularly in the context of virus yellows (BMYV and BChV) in sugar beet. UAVs equipped with high-resolution cameras and advanced image analysis techniques offer several compelling benefits including efficiency, objectivity, cost-effectiveness and consistency (Mahlein et al., 2019). However, it is essential to acknowledge that the complete replacement of human raters with UAVs may not be straightforward and requires careful consideration. In fact, our best model had a RMSE score of 11.45%, indicating that disease incidence can be scored within a confidence interval of ±11.45%. This means that genotypes that differ from each other by a score less than 11.45% disease incidence cannot be accurately scored with the developed UAV-based approach. This limitation is particularly significant when dealing with genotypes that exhibit subtle variations in disease susceptibility.

Complete replacement of human raters by UAVs should be approached with caution, especially in situations where precise discrimination between closely related disease states or genotypes is paramount. Instead, a more pragmatic approach involves leveraging UAVs to streamline the assessment process, improve efficiency and reduce human subjectivity while preserving the role of human experts to validate results, provide domain-specific insights and address cases that fall within the margin of error. In summary, our findings demonstrate the potential of UAVs as powerful tools for disease phenotyping. UAVs equipped with high-resolution cameras can swiftly and objectively assess disease incidence, thereby mitigating the subjectivity and resource-intensive nature of traditional human scoring. The development and application of CNN models further enhance the accuracy and efficiency of disease scoring, offering the promise of rapid advancements in sugar beet breeding. Nevertheless, the path forward involves a nuanced approach, as technology should complement rather than entirely replace human expertise. By harnessing the strengths of both, the phenotyping of disease-resistant sugar beet cultivars can be expedited and optimized.

## ACKNOWLEDGEMENTS

N.O. is grateful to the NGO Förderverein Uni Kinshasa e. V., fUNIKIN, Else-Kroener-Fresenius Stiftung and Holger-Poehlmann Foundation for their support through the excellence scholarship programme 'Bourse d'Excellence Bringmann aux Université Congolaises (BEBUC)'. This work has partly been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC 2070–390732324 and by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the parliament of the federal republic of Germany via the federal office for agriculture and food (BLE) under the strategy for digitalisation in agriculture FKZ 28DE104A18. The authors are also thankful to the technicians who supported image collection and field scoring.

## CONFLICT OF INTEREST STATEMENT

There is no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

## ORCID

*Nathan Okole* https://orcid.org/0000-0002-5375-2420
*Roxana Hossain* https://orcid.org/0000-0003-3536-0957

## REFERENCES

Aasen, H., Honkavaara, E., Lucieer, A. & Zarco-Tejada, P.J. (2018) Quantitative remote sensing at ultra-high resolution with UAV

spectroscopy: a review of sensor technology, measurement procedures, and data correction workflows. *Remote Sensing*, 10, 1091.

Agisoft LLC. (2022) *Agisoft Metashape User Manual: Professional Edition, Version 1.8*. Available from: https://www.agisoft.com/pdf/metashape-pro_1_8_en.pdf [Accessed 2nd July 2024].

Alves, K.S. & Del Ponte, E.M. (2021) Analysis and simulation of plant disease progress curves in R: introducing the epifitter package. *Phytopathology Research*, 3, 22.

Barreto, A., Ispizua Yamati, F.R., Varrelmann, M., Paulus, S. & Mahlein, A.-K. (2023) Disease incidence and severity of Cercospora leaf spot in sugar beet assessed by multispectral unmanned aerial images and machine learning. *Plant Disease*, 107, 188–200.

Bilbrey, J.A., Heindel, J.P., Schram, M., Bandyopadhyay, P., Xantheas, S.S. & Choudhury, S. (2020) A look inside the black box: using graph-theoretical descriptors to interpret a continuous-filter convolutional neural network (CF-CNN) trained on the global and local minimum energy structures of neutral water clusters. *The Journal of Chemical Physics*, 153, 24302.

Bock, C.H., Barbedo, J.G.A., Del Ponte, E.M., Bohnenkamp, D. & Mahlein, A.-K. (2020) From visual estimates to fully automated sensor-based measurements of plant disease severity: status and challenges for improving accuracy. *Phytopathology Research*, 2, 9.

Bock, C.H., Chiang, K.-S. & Del Ponte, E.M. (2022) Plant disease severity estimated visually: a century of research, best practices, and opportunities for improving methods and practices to maximize accuracy. *Tropical Plant Pathology*, 47, 25–42.

Chivasa, W., Mutanga, O. & Biradar, C. (2020) UAV-based multispectral phenotyping for disease resistance to accelerate crop improvement under changing climate conditions. *Remote Sensing*, 12, 2445.

De Mendiburu, F. (2014) Agricolae: statistical procedures for agricultural research. *R Package Version, 1*. Available from: https://cran.r-project.org/web/packages/agricolae/agricolae.pdf [Accessed 2nd July 2024].

Dhaka, V.S., Meena, S.V., Rani, G., Sinwar, D., Ijaz, M.F. & Woźniak, M. (2021) A survey of deep convolutional neural networks applied for prediction of plant leaf diseases. *Sensors*, 21, 4749.

Gowda, S.N. & Yuan, C. (2019) ColorNet: investigating the importance of color spaces for image classification. In: Jawahar, C.V., Li, H., Mori, G. & Schindler, K. (Eds.) *Computer vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, revised selected papers, part IV 14*. Lecture Notes in Computer Science vol. 11364. Cham: Springer, pp. 581–596.

Günder, M., Ispizua Yamati, F.R., Kierdorf, J., Roscher, R., Mahlein, A.K. & Bauckhage, C. (2022) Agricultural plant cataloging and establishment of a data framework from UAV-based crop images by computer vision. *GigaScience*, 11, giac054.

Hossain, R., Menzel, W., Lachmann, C. & Varrelmann, M. (2021) New insights into virus yellows distribution in Europe and effects of beet yellows virus, beet mild yellowing virus, and beet chlorosis virus on sugar beet yield following field inoculation. *Plant Pathology*, 70, 584–593.

Ispizua Yamati, F.R., Günder, M., Barreto Alcántara, A.A., Bömer, J., Laufer, D., Bauckhage, C. et al. (2024) Automatic scoring of Rhizoctonia crown and root rot affected sugar beet fields from orthorectified UAV images using machine learning. *Plant Disease*, 108, 711–724.

Justice, C.O. & Townshend, J.R.G. (1981) Integrating ground data with remote sensing. In: Townshend, J.R.G. (Ed.) *Terrain analysis and remote sensing*. London: George Allen & Unwin, pp. 38–58.

Krishnaswamy Rangarajan, A. & Purushothaman, R. (2020) Disease classification in eggplant using pre-trained VGG16 and MSVM. *Scientific Reports*, 10, 2322.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2017) ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 84–90.

Li, Y., Nie, J. & Chao, X. (2020) Do we really need deep CNN for plant diseases identification? *Computers and Electronics in Agriculture*, 178, 105803.

Lin, L.I.-K. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255–268.

Lin, J., Chen, X., Pan, R., Cao, T., Cai, J., Chen, Y. et al. (2022) GrapeNet: a lightweight convolutional neural network model for identification of grape leaf diseases. *Agriculture*, 12, 887.

Liu, Q., Basu, S., Ganguly, S., Mukhopadhyay, S., DiBiano, R., Karki, M. et al. (2020) Deepsat v2: feature augmented convolutional neural nets for satellite image classification. *Remote Sensing Letters*, 11, 156–165.

Mahlein, A.-K., Kuska, M.T., Thomas, S., Wahabzada, M., Behmann, J., Rascher, U. et al. (2019) Quantitative and qualitative phenotyping of disease resistance of crops by hyperspectral sensors: seamless interlocking of phytopathology, sensors, and machine learning is needed! *Current Opinion in Plant Biology*, 50, 156–162.

Malebary, S.J. & Khan, Y.D. (2021) Evaluating machine learning methodologies for identification of cancer driver genes. *Scientific Reports*, 11, 12281.

Meier, U., Bleiholder, H., Buhr, L., Feller, C., Hack, H., Heß, M. et al. (2009) The BBCH system to coding the phenological growth stages of plants—history and publications. *Journal für Kulturpflanzen*, 61, 41–52.

Muneer, A. & Fati, S.M. (2020) Efficient and automated herbs classification approach based on shape and texture features using deep learning. *IEEE Access*, 8, 196747–196764.

Purkis, S.J. & Klemas, V.V. (2011) *Remote sensing and global environmental change*. Oxford: John Wiley & Sons.

Python Software Foundation. (2021) *Python 3.9.2 Documentation*. Python Software Foundation. Available from: https://docs.python.org/3.9/ [Accessed 2nd July 2024].

R Core Team. (2016) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.r-project.org/ [Accessed 2nd July 2024].

Ras, G., Xie, N., Van Gerven, M. & Doran, D. (2022) Explainable deep learning: a field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73, 329–396.

Saleem, M.H., Potgieter, J. & Arif, K.M. (2019) Plant disease detection and classification by deep learning. *Plants*, 8, 32–34.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X. et al. (2020) Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2, 476–486.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, pp. 618–626.

Soori, M., Arezoo, B. & Dastres, R. (2023) Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 3, 54–70.

Tasdizen, T., Sajjadi, M., Javanmardi, M. & Ramesh, N. (2018) Improving the robustness of convolutional networks to appearance variability in biomedical images. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Washington, DC: IEEE, pp. 549–553.

Thanapol, P., Lavangnananda, K., Bouvry, P., Pinel, F. & Leprévost, F. (2020) Reducing overfitting and improving generalization in training convolutional neural network (CNN) under limited sample sizes in image recognition. In: *2020-5th International Conference on Information Technology (InCIT)*. Chonburi, Thailand: IEEE, pp. 300–305.

Therrien, R. & Doyle, S. (2018) Role of training data variability on classifier performance and generalizability. In: *Medical Imaging 2018: Digital Pathology*, Vol. 10581. Texas, United States: SPIE.

Trapp, J.J. (2015) *Genetics of drought tolerance in common bean* (Phaseolus vulgaris *L.*). [PhD thesis]. Washington, DC: Washington State University.

Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A.J. (2019) Machine learning algorithm validation with a limited sample size. *PLoS One*, 14, e0224365.

Wong, S.C., Gatt, A., Stamatescu, V. & McDonnell, M.D. (2016) Understanding data augmentation for classification: when to warp? In: *2016 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016*. Gold Coast, QLD, Australia: IEEE.

Yang, G., Liu, J., Zhao, C., Li, Z.Z., Huang, Y., Yu, H. et al. (2017) Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. *Frontiers in Plant Science*, 8, 1111.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Okole, N., Ispizua Yamati, F.R., Hossain, R., Varrelmann, M., Mahlein, A.-K. & Heim, R.H.J. (2024) Aerial low-altitude remote sensing and deep learning for in-field disease incidence scoring of virus yellows in sugar beet. *Plant Pathology*, 00, 1–15. Available from: https://doi.org/10.1111/ppa.13973