Affordable High Throughput Field Detection of Wheat Stripe Rust Using Deep Learning with Semi-Automated Image Labeling

Zhou Tang[1], Meinan Wang[2], Michael Schirrmann[3], Karl-Heinz Dammer[3], Xianran Li[4], Robert Brueggeman[1], Sindhuja Sankaran [5], Arron H. Carter[1], Michael O. Pumphrey[1], Yang Hu[1*], Xianming Chen[2,4*], and Zhiwu Zhang[1*]

[1]Department of Crop and Soil Sciences, Washington State University, Pullman, WA 99164, USA
[2]Department of Plant Pathology, Washington State University, Pullman, WA 99164, USA
[3]Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB), Potsdam-Bornim, Max-Eyth-Allee 100, 14469, Potsdam, Germany
[4]USDA-ARS Wheat Health, Genetics, and Quality Research Unit, Pullman, WA 99164, USA
[5]Department of Biological Systems Engineering, Washington State University, Pullman, WA 99164, USA

[*]Corresponding should be addressed to YH (yang.hu@wsu.edu), XC (xianming.chen@usda.gov), or ZZ(zhiwu.zhang@wsu.edu)

Abstract

Stripe rust (caused by *Puccinia striiformis* f. sp. *tritici*) is one of the most devastating diseases of wheat and causes large-scale epidemics and severe yield loss. Applying fungicides during early epidemic development is crucial to controlling the disease but is often challenged by resource-limited human visual scouting. Deep learning has the potential to process images and videos captured from affordable devices to empower high-throughput phenotyping for early detection of stripe rust for timely application of fungicides and improve control efficiency. Here, we developed RustNet, a neural network-based image classifier, for efficiently monitoring fields for stripe rust. RustNet was built on a ResNet-18 architecture pre-trained with ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) dataset using transfer learning. RGB images and videos of multiple wheat fields with different wheat types (winter and spring wheat), conditions (irrigated and non-irrigated), and locations were acquired using smartphones or unmanned aerial vehicles near the canopy. A semi-automated image labeling approach was conducted to improve labeling efficiency by combining automated machine labeling and human correction. Cross-validations across multiple categories (sensor platforms, wheat types, and locations) achieved Area Under Curve from 0.72 to 0.87. Independent validation on a published dataset from Germany achieved accuracies ranging from 0.79 to 0.86. The visualization of the last convolutional layer of RustNet demonstrated the identification of pixels with stripe rust. RustNet is freely available at https://zzlab.net/RustNet.

Keywords: Plant disease, Machine vision, UAV, Smartphone, Convolutional Neural Network

Introduction

Wheat (*Triticum aestivum* L.) is grown in more areas than any other crop and is the leading source of plant-based protein in the human diet (Salamini et al., 2002; Shiferaw et al., 2013). Wheat production is threatened by plant pathogens and pests, which annually cause around 20% yield loss in wheat globally (Savary et al., 2019). Wheat stripe rust (caused by *Puccinia striiformis* f. sp. *tritici*) is a devastating airborne fungal disease and occurs worldwide in all major wheat-growing regions (X. Chen, 2020; X. M. Chen, 2005). Stripe rust is a polycyclic disease with numerous infections occurring in a growing season. Urediniospores of the pathogen travel thousands of kilometers by wind and infect any above-ground wheat organs throughout all growth stages of susceptible wheat cultivars whenever weather conditions are favorable for infection (X. M. Chen, 2005). The pathogen is capable of causing large-scale epidemics and has the potential to cause 100% yield losses, which significantly threatens global food security (X. Chen, 2020).

The development of resistant cultivars and the application of fungicides are major approaches to controlling stripe rust (X. Chen, 2013). Suppose resistant cultivars become susceptible due to the development of new virulent races (Liu et al., 2017), or susceptible cultivars are grown for other reasons. In that case, fungicide application can be used to prevent epidemics or reduce yield losses. When fungicide applications are necessary, timing is critical for the application. If the application is too early, the effective fungicide period cannot cover the period of pathogen infection and disease progress. The fungicide cannot effectively suppress disease development (Kang et al., 2019). Both situations ultimately may lead to failure in controlling the disease. Currently, timely decision-making to apply fungicides requires plant pathologists and growers to scout the fields and monitor pathogen presence visually. Although detecting the incidence of stripe rust has been conducted through human observation for decades (Wang et al., 2022), this process is laborious and inefficient, especially for large production fields. Therefore, developing an efficient way to monitor stripe rust at low infection rates across large production fields is necessary.

Unmanned aerial vehicles (UAVs) have been tested with multispectral and hyperspectral sensors for wheat stripe rust detection. The pixels of UAV-based multispectral imagery can be classified into healthy and infected wheat plants via a random forest classifier with Ratio Vegetation Index (RVI), Normalized Difference Vegetation Index (NDVI), and Optimized Soil Adjusted Vegetation Index (OSAVI) (Su et al., 2018). For hyperspectral imagery, previous research demonstrated that Deep Learning (DL) had higher accuracy (0.85) than a conventional Machine Learning (ML) algorithm (random forest, 0.77) to detect stripe rust using manually labeled UAV images region-wisely (X. Zhang et al., 2019). Additional research also confirmed that DL outperformed the random forest method based on images with pixel-wise labels (Su et al., 2021). However, information on the spatial patterns of the disease within the fields is needed to accurately coordinate the application of fungicides over time and space, at best with low-cost sensors (Oerke, 2020).

RGB (Red, Green, and Blue) cameras have been exploited as an affordable option for monitoring crop diseases to overcome the high cost of multispectral and hyperspectral cameras. Smartphones were investigated to detect wheat stripe rust using RGB images on a single leaf (Mi et al., 2020). Experts took wheat stripe rust images with hand-held smartphones and cropped the images containing only one leaf. They labeled the leaves into five infection types of wheat stripe rust to train a convolutional neural network (CNN) (Mi et al., 2020). Winter wheat fields were imaged with a digital single-lens mirrorless (DSLM) camera installed on a movable boom arm and trained a deep residual neural network to distinguish between infected and healthy leaves (Schirrmann et al., 2021). However, collecting RGB images with these two methods (smartphone and DSLM camera) was not as efficient as the UAV-based RGB imagery. UAVs can automatically fly in the field and efficiently cover a large area capturing high-resolution imagery (C. Zhang & Kovacs, 2012).

A large image dataset with accurate labels is the key to training a classifier under a supervised learning paradigm. Big public image datasets, such as ImageNet labeled over 14 million images from over 21 thousand classes (Russakovsky et al., 2015), has significantly promoted the progress of computer vision. In 2015, the residual neural network (ResNet) achieved 3.57% on the top-5 error (measure if model correct among top 5 predictions) in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) which contains over 1.4 million labeled images from 1,000 object classes (He et al., 2015). ResNet can stack different numbers of layers. For example, the ResNet-18 is named after its total trainable 18 layers in the structure. The key technology of ResNet is the identity shortcut connection that can skip several convolutional layers to overcome the vanishing gradient problem and makes it possible to efficiently train a deep neural network of over 1,000 layers (He et al., 2015).

Existing technologies have the potential to detect wheat stripe rust in an affordable high throughput fashion except for two challenges. One is to efficiently collect field images at an affordable cost. The other is to develop a platform to efficiently label images to train a deep neural network. In this study, we collected RGB images with UAV and smartphones over heterogeneous wheat fields with different wheat types (winter and spring wheat) and conditions (irrigated and non-irrigated). Our objectives were to 1) implement a semi-automated image labeling strategy to efficiently label wheat stripe rust images; 2) evaluate the ResNet-18 ability for wheat stripe rust detection using cross-validations across datasets; and 3) release a ResNet-based neural network model (RustNet) to incorporate Rooster software (https://zzlab.net/Rooster) so that users can import a large amount of field images for the detection of wheat stripe rust.

Results

Semi-automatic image labeling during model development

We manually labeled 56 images to validate the development of RustNet at different stages. These images only contained partially diseased leaves. RustNet achieved an AUC of 0.64 (**Figure 4**) after training RestNet18 with images that were labeled automatically (Stage 1), compared with an AUC of 0.5 with random guesses. The result suggests that the abundantly available images without diseased leaves and images with all plants diseased are valuable to developing RustNet with automatically labeling to detect stripe rust. With the prediction guidance from Stage-1 RustNet, human adjustments on images with all plants diseased boosted RustNet into stage 2 with an AUC of 0.75. With the prediction guidance from Stage-2 RustNet, human adjustments on images with partially infected leaves enhanced RustNet into stage 3 with an AUC of 0.85.

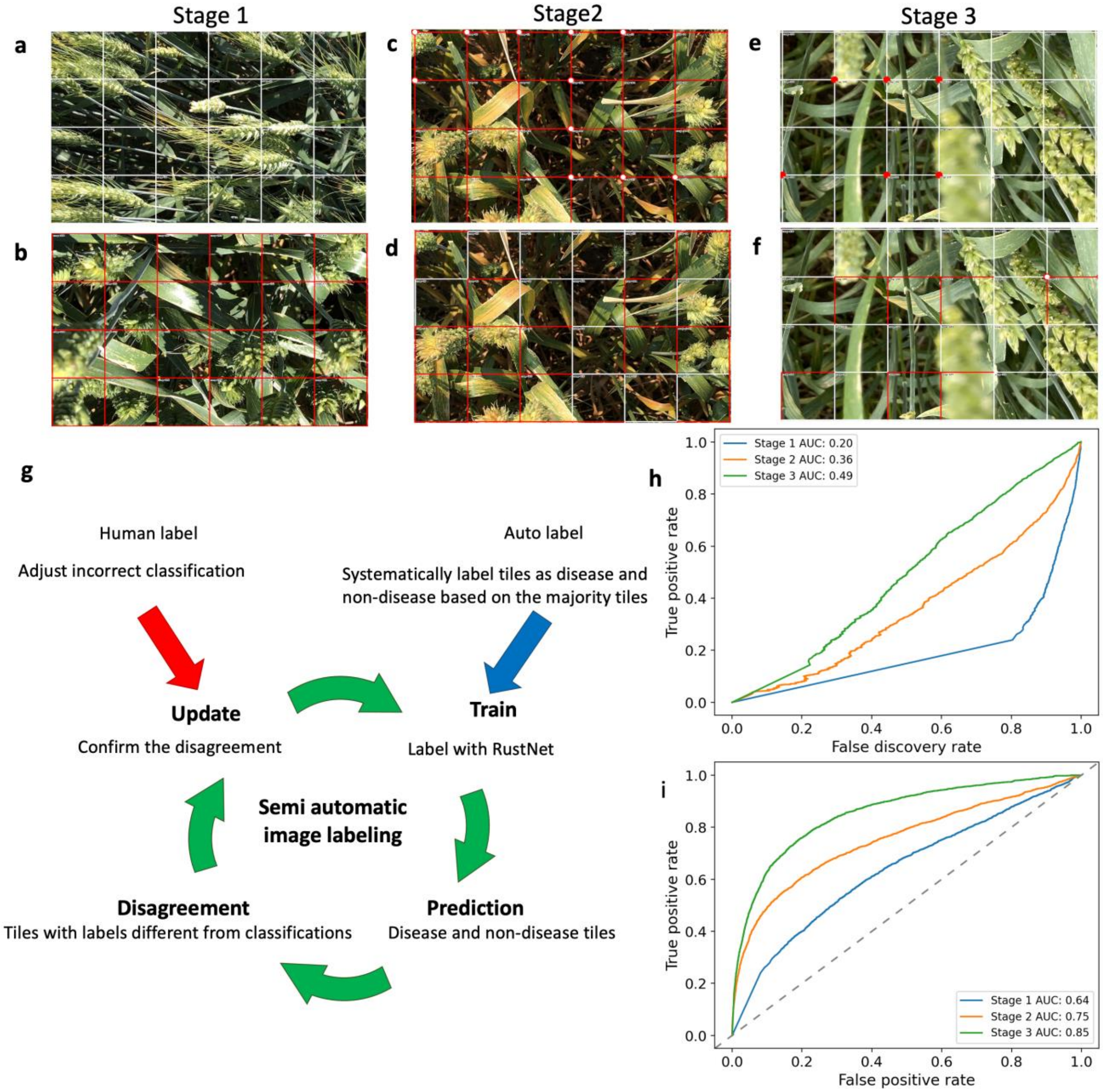


Figure 4. A pipeline of using ROOSTER and RestNet18 to develop RustNet using semi-automatic image labeling. Labeling software ROOSTER allows users to define tiles on an image with the number of rows and columns. The default statuses are non-disease (a) which can be changed to disease by clicking the Reverse button (b) for all tiles or double-clicking a specific tile. The two types of images with no disease and all plants infected were automatically labeled (Stage 1) to train RestNet18 to develop RustNet. At Stage 2, different images with all plants infected are initiated with disease status and predicted by RustNet (c). The dots indicate disagreement between predictions and current statuses. The predictions navigate humans to adjust labels (d) to further train RustNet. At Stage 3, images with partially infected leaves are initiated with non-disease status and predicted by RustNet (e). The predictions navigate humans to adjust labels (f) to further train RustNet. The three stages were repeated for all available images (g). The performance of RustNet

at different stages was examined by 61 images with partial leaves diseased and all tiles labeled manually. The performances are indicated by the receiver operating characteristic curves with true positive rate against false discovery rate (h) and false positive rate (i). A diagonal dash line was added to the ROC with false positives rate, showing an AUC of 0.5 for the random guess (i) compared to 0 in the ROC with false discovery rate (h).

Similarly, with the 20,360 published tile images (5,818 diseased tiles and 14,542 non-diseased tiles) from Marquardt, the AUCs are 0.64, 0.78, and 0.87, for RustNet at Stage 1, 2, and 3, respectively (**Table S2**). Again, the AUC of 0.64 from Stage-1 RustNet validated the value of automatic labels on the abundantly available images without diseased leaves and images with all plants diseased. Similar performance of Stage-3 RustNet can also be achieved by thoroughly training RestNet18 with six sets of images labeled during the development of RustNet. The median AUC is 0.855 with a minimum of 0.78 and a maximum of 0.87. Semi-automatic image labeling can improve classification accuracy step by step.

Classification validations within images collected in Pullman

Two types of validations were conducted within images collected in Pullman. The first classification validation was conducted under the scheme of image collection platforms (images from UAV, images from smartphone, and images extracted from video by UAV and smartphone), which split images into three cohorts by platforms. Two cohorts were used as training data and tested in the third cohort. These AUCs (**Figure 5a-b**) were almost the same when testing with images from different collection platforms (AUC of 0.79, 0.80, and 0.79 for phone, video, and UAV separately). However, RustNet obtained varied FDRs for the same TPR when testing images from different platforms, especially testing with still images of UAVs. This difference suggested there was a lower prevalence of disease in UAV images than in other sources.

The second classification validation was conducted under the scheme of wheat types (spring wheat in Spillman farm and winter wheat in PCFS). Images of one wheat type were used as training data and tested using images of another wheat type (**Figure 5c-d**). Consistently, training RustNet with images of spring wheat achieved higher AUC (0.76) than training with images of winter wheat (0.72). The drought conditions may have caused difficulty to differentiate diseased and non-diseased leaves in the winter wheat trials.

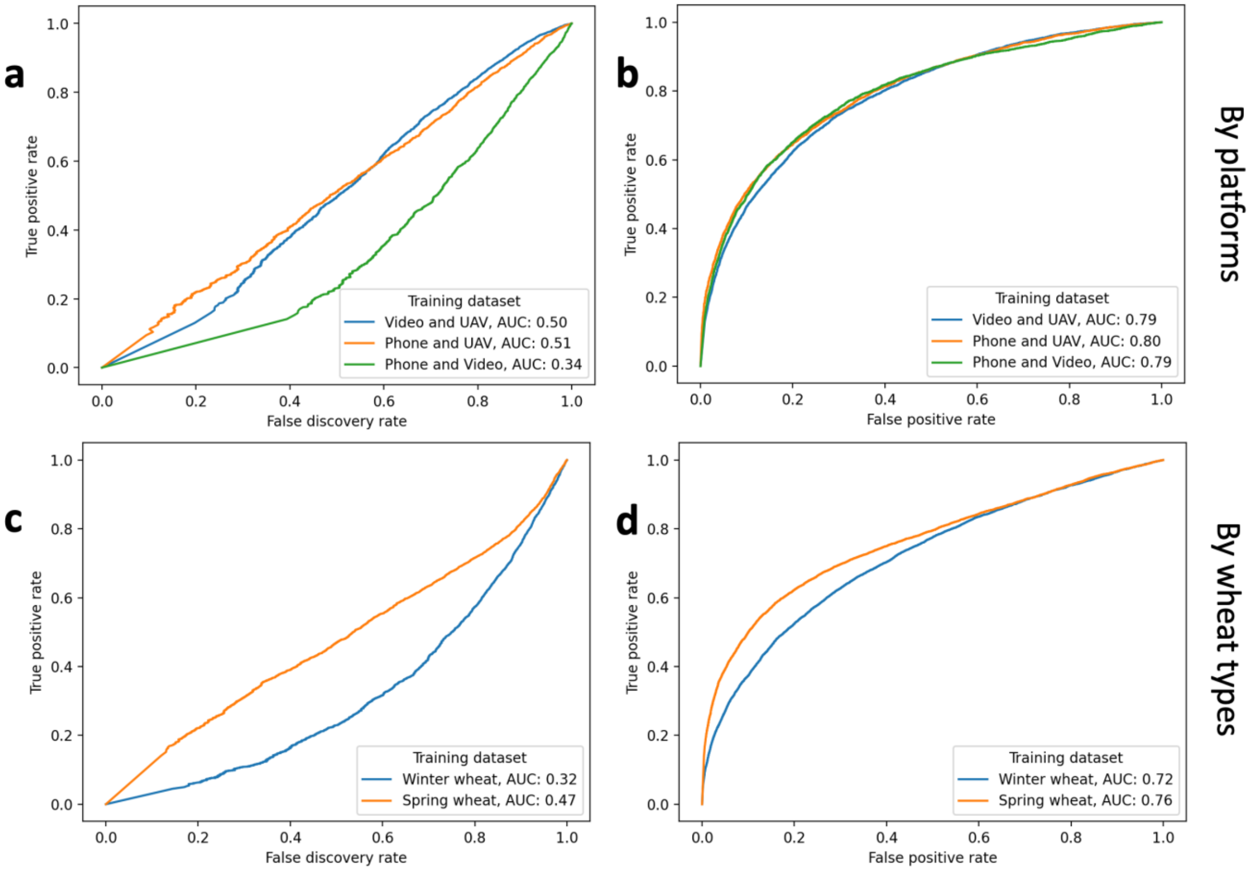


Figure 5. Cohort validation across platforms and wheat types within Pullman. When the image data were split based on platforms scheme (still phone image, still UAV images, and images from videos by phone of UAV), two of the three platforms were used as training data to test the other one. The model performances were investigated with the true positive rate under different levels of (a) false discovery rate and (b) false positive rate. When the image data were split based on the wheat types (spring wheat or winter wheat), one wheat type was used as training data and tested using images of another wheat type. The model performances were investigated with the true positive rate under different levels of (c) false discovery rate and (d) false positive rate.

Independent validation across locations

RustNet was developed through interactive deep learning that integrated labeling images collected in Pullman, WA, USA, under the ResNet-18 backbone. The predictions of RustNet were used as the initial image label, followed by human adjustments. To test if RustNet would work in other images, we tested the predictions on the previously published images collected in Marquardt, Potsdam, Germany, labeled independently by different experts. The accuracy trend curves for the training process are shown in **Figure S1**. The majority of prediction scores for tiles in different DAI were distributed near zero or one (**Figure S2**) because of the softmax function in the last layer of the neural network. RustNet developed using the images collected from Pullman worked well (AUC with FDR = 0.68) to predict the images contained from Marquardt (**Figure 6a**).

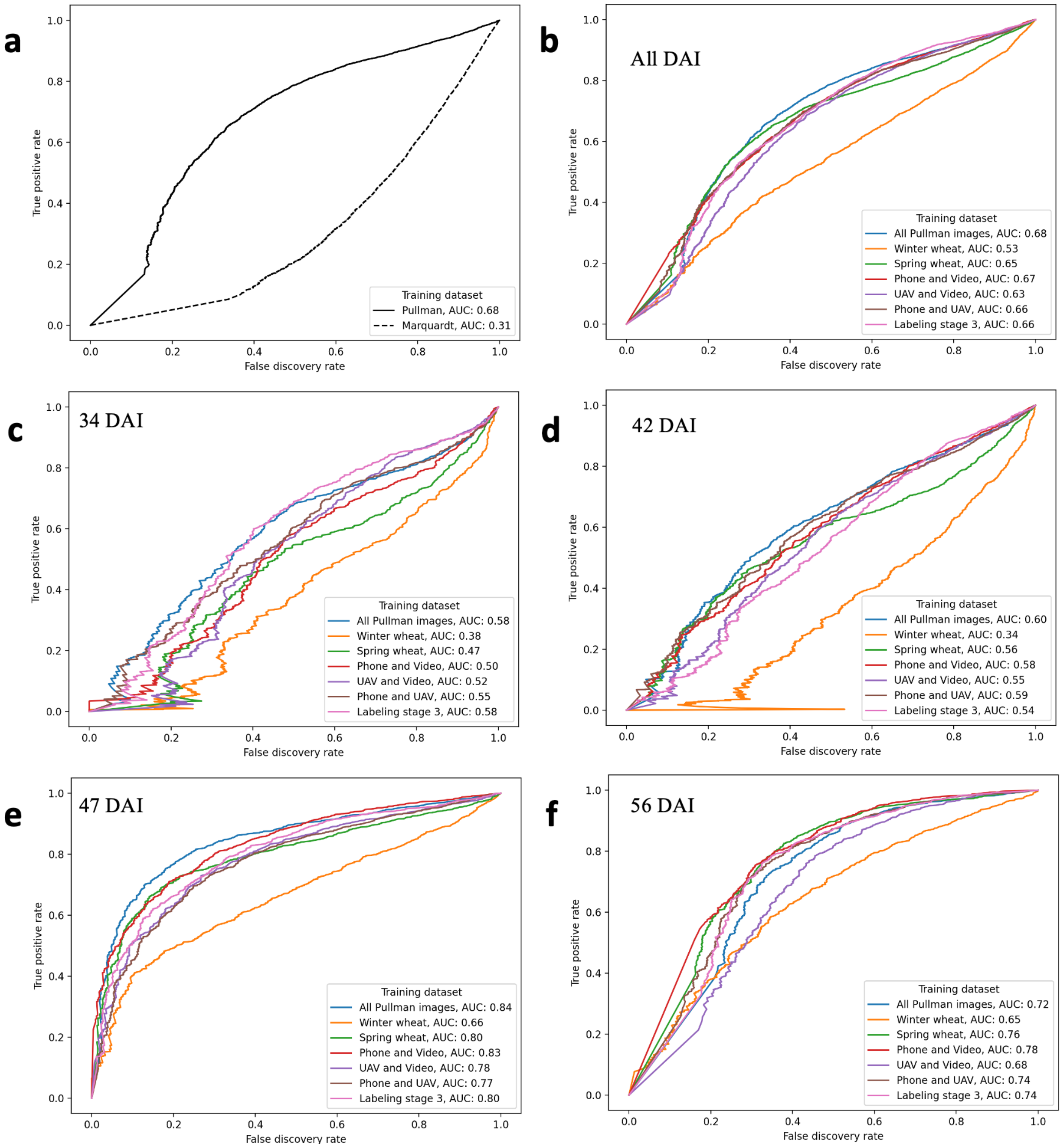


Figure 6. Classification validation across locations. Model performances were investigated with the true positive rate under different levels of false discovery rate. Images were split into two parts based on the location scheme (Pullman, WA, USA and Marquardt, Potsdam, Germany). Images from one location were used as the training data and tested using another location (a). Seven

parameter sets were trained with different parts of the Pullman data and tested with images from Marquardt on (b) all DAI, (c) 34 DAI, (d) 42 DAI, (e) 47 DAI and (f) 56 DAI (days after inoculation, DAI)

We also did a reverse independent validation by retraining RestNet18 with Marquardt images and testing the prediction on the Pullman images (**Figure 6a**, the dashed line). Because the Pullman images were more diversified, containing two wheat types (spring and winter), two conditions (irrigated and non-irrigated), multiple platforms (images taken by UAV, images taken by smartphone, and images extracted from video), and different resolutions, as expected, the reverse validation resulted in poor performance. The AUC with FDR of reverse validation was reduced to 0.31, compared to 0.68 in the other direction.

The validations were further conducted by using different subsets of training from Pullman and different subsets of testing data from Marquardt. The training subsets included 1) the 105 images from Stage 3 in developing RustNet; 2) spring wheat; 3) winter wheat; 4) phone and video; 5) UAV and video; 6) phone and UAV, and 7) all images from Pullman. The testing datasets included Marquardt images from all DAI (**Figure 6b**) and images from specific DAI: 34, 42, 47, or 56 (**Figure 6c** to **f**). In general, using all the data from Pullman performed the best to predict the Marquardt data. Similar to the finding in the validation within the Pullman data, the images of winter wheat had lower accuracy than the images of spring wheat for training the model and testing in the Marquardt data. The AUC was summarized for these training datasets and the training images in the three stages of developing RustNet (**Table S2**). It is interesting to note that the 105 images in Stage 3 of developing RustNet almost had a similar accuracy as all the images (300) because the 105 images had complete coverage of all the types of the images.

Various criteria were also evaluated for the different training datasets from Pullman to test images from Marquardt. These criteria included Sensitivity, Specificity, Accuracy, and F1 score (**Table 2**). The threshold was set to 0.5 for the final scores from the last layers of ResNet18 to determine uninfected (≤0.5) and infected (>0.5) images. The specificity and sensitivity were 0.51 and 0.96, respectively, for all DAI. Specificity decreased with DAI as the disease became more severe, while sensitivity had the opposite trend. Accuracy stayed the same across different DAI (0.83-0.86). The exception was 42 DAI with an accuracy of 0.79 when the number of non-disease tiles was the highest across DAI.

Table 2. Classification of tiles of Marquardt, Potsdam, Germany (*n*=20360) into disease and non-disease* classification.

DAI	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy	F1 score
34	227	3,128	35	623	0.27	0.99	0.84	0.41

42	588	4,027	88	1,147	0.34	0.98	0.79	0.49
47	1,083	3,906	58	769	0.58	0.99	0.86	0.72
56	1,088	2,865	435	293	0.79	0.87	0.84	0.75
All	2,986	13,926	616	2,832	0.51	0.96	0.83	0.63

*TP (true positive), TN (true negative), FP (false positive) and FN (false negative) were counted within different days after inoculation (DAI) and all days separately. The RustNet was trained with all images from Pullman. Criteria of 0.5 was used to decide disease or non-disease classification.

Enhancement of negative control

To explore classification ability on images beyond the wheat stripe rust, RustNet was tested with the validation set of ILSVRC (ImageNet Large-Scale Visual Recognition Challenge), which has 50,000 images from 1,000 classes (Russakovsky et al., 2015). After training with all Pullman images, RustNet made diseased predictions on 1,507 images from 651 classes, and the false-positive rate was 3.01%. We randomly selected 80% of images (1,205 out of 1,507) into the training data of the non-disease class to improve negative control. RustNet was re-trained with the updated training dataset and tested with the remaining 20%. The false rate was reduced to 1.66% (5 out of 302 images). The improvement in false positives on the additional images did not affect the performance of testing stripe rust. Compared to the stage before adding negative control images into the training data (**Figure S3**), the updated RustNet did not show much difference in AUC when models were tested separately with the Marquardt's data.

Similarly, we enhanced RustNet with the images from a corn leaf disease dataset (J & GOPAL, 2019; Singh et al., 2020). There were 4,161 images from four classes in the corn leaf disease dataset: corn rust (1,306), corn gray leaf spot (547), corn blight (1,146), and healthy corn leaf (1,162). Before adding these images into the training set as a negative control, the false positive rate was high, 77.11%, 89.76%, 67.89%, and 0.86% for corn rust, gray corn leaf, and corn blight and healthy corn leaf, respectively. We randomly selected 80% false-positive images (1,831 out of 2,288) and added them to the training data for a negative control to update RustNet. When tested with the remaining 20% images (457 out of 2,288), none of them were predicted as diseased, and the false positive rate was zero. The updated RustNet did not show much difference in AUC when tested with Marquardt's data (**Figure S3**).

The highlight of stripe rust at the last layer of RustNet

The Grad-CAM method was applied to visualize the last convolutional layer of RustNet to understand how RustNet that was trained with all Pullman images made predictions about the input images. Since the Grad-CAM is class-specific (Selvaraju et al., 2020), it would highlight the region in a tile image triggering the disease prediction. Tiles with correct disease and non-disease prediction in Marquardt's dataset (**Figure 7**) were randomly selected from different DAI to test RustNet. In the disease tiles, these pixels in red regions were weighted most to activate RustNet to make a disease prediction, whereas blue regions were not indicated as diseased, which overlapped with human observations about where stripe rust was present. In non-disease tiles, none of the pixels were highlighted to activate the neural network to make a disease prediction.

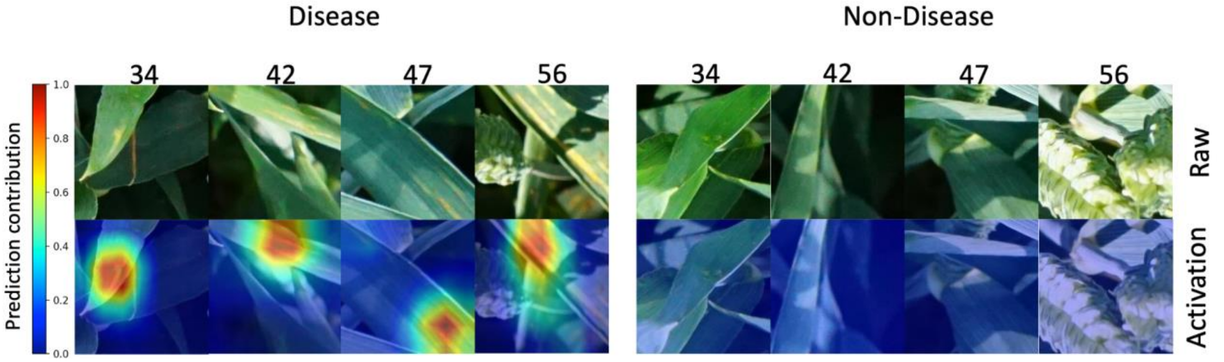


Figure 7. Visualization of RustNet. Row images were captured on different days (34, 42, 47 and 56 days after inoculation) in Marquardt, Potsdam, Germany. Corresponding to each image, the prediction contribution in an activation map highlighted important pixels for a disease prediction. Pixels were more important if the color was closer to red for a disease prediction.

The performances of RustNet trained with all Pullman images were explored on the prediction in Marquardt's dataset from different DAI. In non-disease tiles with disease predictions (**Figure S4**), the activation map captured areas with similar symptoms of wheat stripe rust for a disease prediction. These symptoms can be from drought or other wheat leaf diseases to cause false-positive predictions. In disease tiles with non-disease predictions (**Figure S5**), the wheat stripe rust symptoms were difficult to observe and allowed the RustNet to capture major plant areas for false-negative predictions.

Discussion

Semi-automated image labeling boosts image labeling efficiency and accuracy

As a cyclical re-supervision approach, the semi-automated image labeling can improve labeling efficiency and accuracy. The model that was trained in the previous stage can predict new images from the next stage before laborious human input. The process reduces the human labeling workload. The Rooster software enables the labeling of disease tiles by clicking a mouse. Since classification models can partially predict tiles correctly, people only need to validate the labels.

As the model's accuracy improves, the number of tiles needing human input for label confirmation decreases. This strategy also improves the quality of image labeling. The human labeling can focus on the tiles that were not predicted correctly. With the Rooster software, neighbor tiles can also assist humans in making the right decisions about label classes. When wheat leaves with stripe rust across multiple tiles, labeling these tiles together should be more accurate than labeling a single tile at one time.

The cyclical process benefits the determination if additional labels are needed for a certain type of images. This interactive image annotation can timely visualize images that should be worthy of adding to future training datasets. For example, if the prediction of the previous classification model is almost correct on a new batch of images, these images are not necessary to be added into future training datasets. Unlike labeling all images by experts at one time before training the classifier (DeChant et al. 2017; Lu et al. 2017; Mi et al. 2020; Mohanty, Hughes, and Salathé 2016; Schirrmann et al. 2021), the semi-automated image labeling is a cyclical re-supervision approach that allows users to keep adjusting datasets and improving model accuracy.

Image diversity and adaptation to future applications

This study had wide coverage of diversity in this study benefits the future applications under diversified environments. The fields were imaged with different platforms (UAV, phone, and video). The smartphones and UAVs are more economically feasible than a multispectral camera (Su et al. 2018), hyperspectral camera (Zhang et al. 2019) or LiDAR sensor (Qiu et al. 2019). The fields were also varied by varieties and growth conditions (irrigation vs. non-irrigation). This explained the relative low prediction accuracy competed to the high accuracy in other studies on wheat stripe rust detection (Su et al. 2018, 2021; Zhang et al. 2019). The high accuracy in other studies were also resulted from isolation of stripe rust such as taking images on single leaf or even removing background (Mi et al. 2020).

The image diversity was further enriched in this study for the negative class. Previous research (Schirrmann et al. 2021) indicated that ResNet-18 could be applied in wheat stripe rust detection and multiple crops disease classification. The performance of the neural network could be further improved by adding a wide range of images out of the intended field experiment's context as a negative control. In our research, after training models with stripe rust and non-rust images on wheat, images from the ILSVRC and corn leaf disease datasets were tested. Those false-positive images were added as non-rust images for negative control. RustNet was re-trained with the updated training dataset, which improved the accuracy of the classification model (**Figure S5**).

Prediction assessment

With the cutoff of the prediction score fixed at 0.5, four prediction criteria exhibited different sensitivities. In general, the performances of the model trained with images from the US improved over DAI to predict images from Germany for all the criteria, including sensitivity, specificity, accuracy, and F1 score (**Table 2**). The exceptions were specificity and accuracy. They were similar across different DAI except for specificity at the DAI of 56 (0.87) and accuracy at the DAI of 42 (0.79). They appeared lower than the rest (specificity >= 0.98 and accuracy >= 0.84). With the varied cutoff of prediction score, ROC of TPR against FPR and TPR against FDR also exhibited different sensitivities (**Figure 5**). The three curves of TPR against FPR were very similar among the three validations across platforms of image collections. The AUC is 0.79, 0.80, and 0.79 to test the images collected by smartphones, videos, and UAVs. However, their differences were revealed by the curves of TPR against FDR. The AUC of testing UAV images appears much lower than the other two validations.

There are two possible reasons for the phenomenon. One is related to image diversity, and the other is related to image resolution. UAVs had larger field coverage than smartphones carried by a human. UAVs randomly selected the locations. However, the locations were intentionally emphasized in areas with no plants infected and the areas with plants infected severely using smartphones. The UAV images had better coverage than the smartphone images. The other reason is that UAV images had better resolution (5472 × 3648) than other platforms. The smartphones have a resolution of 4032 × 3024 for still images and 3840 × 2160 for videos. The UAV video has a resolution the same as the smartphone videos.

Conclusions

Timely detection of wheat stripe rust occurrence is important for timely applying fungicides to control the disease epidemic. Deep learning empowers UAVs and smartphones as an efficient phenotyping platform in wheat stripe rust field detection. In this study, we collected images and videos with affordable UAVs and smartphones in diversified fields with different wheat types and growth conditions (drought and irrigation). Images were labeled with a strategy that combines auto machine-labeling and human adjustment into an iterative cycle. Results showed semi-automatic labeling can boost image labeling and classification accuracy step by step. A ResNet-based wheat stripe rust detection neural network, RustNet, was developed to detect visible wheat stripe rust in the field. Cross validation and independent validation proved RustNet can adapt to complex field conditions. We released RustNet to the public and it is freely available at https://zzlab.net/RustNet. The model developing pipeline has potential to apply to broader crops to accelerate precise protection in crops.

Author contributions

Conceptualization: YH, XC., and ZZ. Data analysis: ZT. Data collection: ZT, ZZ, MS, and KD. Field management: XC, MP, and RB. Data labeling: ZT., MW, YH, and ZZ. Wrote manuscript: ZT and ZZ. All authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

This project was partially supported by the USDA National Institute of Food and Agriculture (Hatch project 1014919, Award #s 2018-70005-28792, 2019-67013-29171, and 2020-67021-32460) and the Washington Grain Commission (Endowment and Award #s 126593 and 134574). The authors thank Wanling Li and Jia Yao for data collection.

Materials and Methods

Plant material and field experiment

Two experimental wheat fields were imaged in Pullman, WA, the US, in 2021. The first field (**Figure 1a**) was at the Palouse Conservation Field Station (PCFS, 46°45'36.2 "N, 117°11'59.3 "W), which consisted of two experimental winter wheat trials and was under a rainfed production system (33 cm of precipitation in 2021, non-irrigated trial). One trial, planted with susceptible winter wheat variety 'PS 279', was used for testing various fungicides, and another trial was used for assessing stripe rust resistance levels and potential yield losses of 23 major commercially grown winter wheat cultivars, in addition to 'PS 279' used as a susceptible check (https://striperust.wsu.edu/). The first trial was a randomized complete block design experiment, and the second was a randomized split-plot design experiment. Both trials were planted on November 1, 2020, with four replications, and each plot measured 1.37 m × 4.88 m, and the borders of experimental plots were surrounded by 'PS 279' to induce uniform infection. The plants in the surrounding borders were inoculated twice with urediniospores of *P. striiformis* f. sp. *tritici* (mainly race PSTv-37) collected from the same location in 2020 to initiate disease. The first inoculation was conducted on April 24, 2021, when most plants were at late tillering (Zadoks GS 26), and the second on June 12, 2021, when most plants were at boot stage (GS 45) (ZADOKS et al., 1974). The first trial was applied with various fungicides scheduled on May 11 at GS 30 and/or May 26 at GS 45, and the second trial was applied with fungicide Quilt Xcel at the commercial rate (14 fl oz/A) on May 11 (GS 30) and May 26 (GS 45) to the plots scheduled for fungicide application. Due to the extremely dry and hot conditions, stripe rust did not develop to an epidemic level, and the original objectives of the trials could not be achieved, but the low rust level was suitable for the purpose of the present study as the negative control.

The second field (**Figure 1b**) consisted of spring wheat nurseries located in the Spillman Agronomy Farm (46°41'41.1 "N, 117°08'30.3 "W), where the regular irrigation was carried out during the crop season to deliver 46 cm of water to the crop. The spring wheat cultivar (Lemhi 66) grown in the borders is highly susceptible to stripe rust. Wheat plants surrounding three sides of the borders (top, left, and bottom) were artificially inoculated with urediniospores of *P. striiformis* f. sp. *tritici*, whereas the right border was not.

Image acquisition

A DJI Air 2s UAV (SZ DJI Technology Co., Shenzhen, China) was used to take static images and videos. The UAV has a camera with a one-inch CMOS (Complementary Metal-Oxide Semiconductor) sensor containing 20 megapixels. The equivalent focal length of the lens in this camera is 22 mm, with an aperture of f/2.8 and a field angle of 88 degrees. The resolution of an RGB image was 5,472 × 3,648 pixels (**Table 1**). The exposure time and ISO (International Organization for Standardization) of the camera were set to the auto model. The UAV was maintained at around 1.2 meters by hovering above the wheat canopy surface when capturing images and videos. Each image captured a roughly 2.7 m × 1.8 m wheat plot area from this altitude, and the ground sampling distance was around 0.5 mm/pixel. Examples of UAV images with diseased and non-diseased leaves are shown in **Figure 1c**. Videos acquired with the UAV were 30 frames per second (fps) at a flight speed of around 2 meters per second. One frame was extracted out of every ten frames for UAV videos to ensure three frames were extracted per second. Blurry images were removed with a python script, and repeat frames were checked manually. The image resolution of each video frame was 3,840 × 2,160 pixels (**Table 1**).

Table 1. Summary of images collected at Pullman, WA, USA for use as the training dataset

Farm, Crop, Condition	Number of raw images	Image size (pixels)	Number of diseased tiles	Number of non-diseased tiles	Proportion of diseased tiles (%)
Spillman, Spring wheat, Irrigation	50 (images by UAV)	5472 × 3648	1,726	17,474	8.99
	50 (images by phone)	4032 × 3024	3,278	10,192	24.34
	50 (video by phone)	3840 × 2160	2,462	5,188	32.18
PCFS, Winter wheat, Rainfed	50 (images by UAV)	5472 × 3648	2,313	16,887	12.05
	50 (images by phone)	4032 × 3024	4,726	7,298	39.30
	50 (video by UAV)	3840 × 2160	1,576	6,924	18.54
Total	300	-	16,081	63,963	20.09

An iPhone 8 (Apple Inc, CA, USA) was held in hand to capture images and videos. The camera of the smartphone has a CMOS sensor with 12 megapixels and an aperture of f/1.8. The resolution of an RGB image from the camera was 4,032 × 3,024 pixels. The exposure time was set to auto mode during photography. An example of the smartphone static image tiles with diseased and non-diseased leaves from two farms is shown in **Figure 1c**. Smartphone videos of 60 fps were also collected, and one image was extracted from every 20 frames, leading to three images being extracted per second. Blurry images were removed with Python script, and repeat frames were checked manually. The image resolution from each frame of the phone videos was 3,840 × 2,160 (**Table 1**).

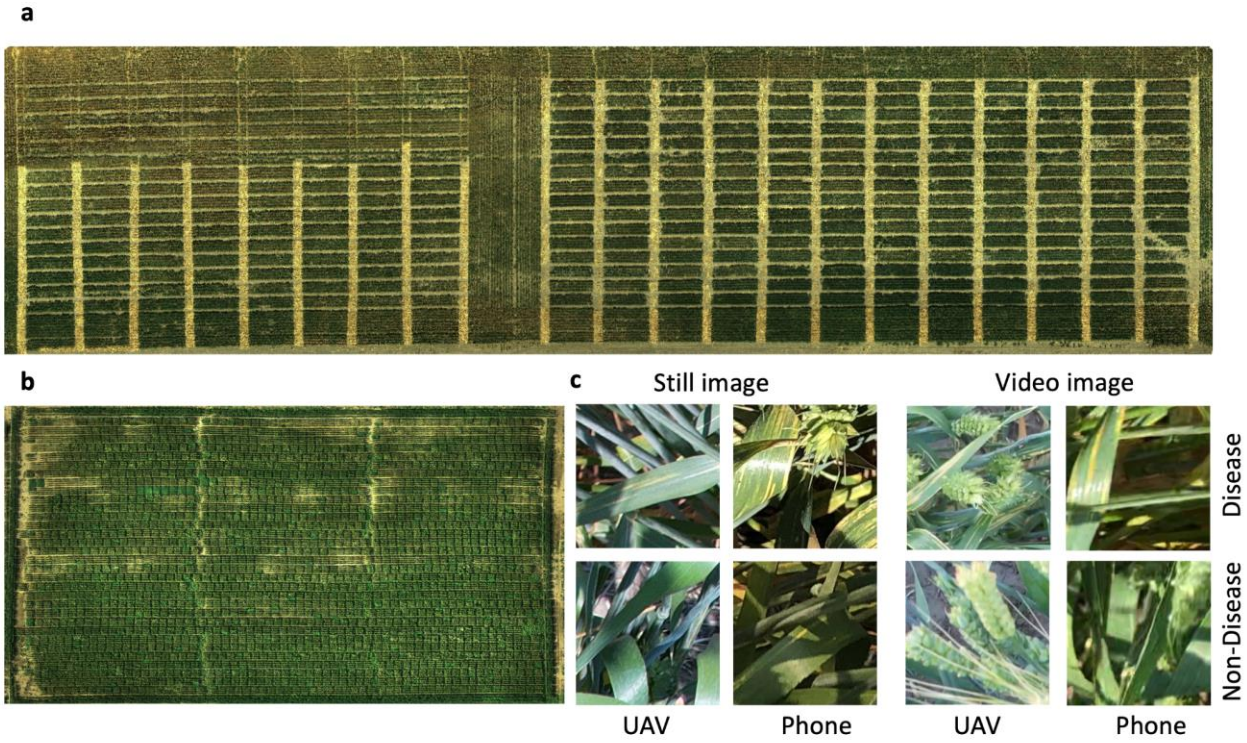


Figure 1. Overview of experimental fields in Pullman and demonstration tile images. Two fields located in PCFS (a) and Spillman farms (b) in Pullman, WA, USA, were surveyed with images and videos taken by a UAV (DJI Air 2S) and a smartphone (iPhone 8). The ortho images over PCFS farm (a) were taken on July 13, 2021, and Spillman farm (b) was taken on July 4, 2021. The still images or images captured from videos were divided into tiles (e.g., 224 × 224 pixels) and labeled as disease or non-disease (c).

A published set of images collected in Marquardt, Potsdam, Germany (Schirrmann et al., 2021) was also used for independent validation. Winter wheat variety Matrix B, which is highly susceptible to stripe rust, was sown in the study. All plots were evenly sprayed with urediniospores, and fungicide (Osiris®, BASF, Germany) was only applied to control plots. Images were collected on four different days (34, 42, 47, and 56 days after-inoculation, DAI) with a digital single-lens mirrorless (DSLM) camera installed on a movable boom arm 2 meters above the ground. This dataset included 5,818 tiles of disease class and 14,542 tiles of non-disease class labeled manually. Raw images number and tiles numbers for infection and non-infection classes on each day were summarized in **Table S1**.

Training and visualization of RustNet

ResNet-18 was selected as the basis of RustNet for image classification. The shortcut connection was implemented as a residual block which suggests identical addition or down-sampling before the addition (**Figure 2a**). The initial parameters of the ResNet-18 were pre-trained with ImageNet data. The output dimension was modified into two classes. Before being fed into the neural network, tile images were resized as 224 × 224 pixels and normalized by converting the range between zero and one. All parameters of the ResNet-18 were retrained for 100 epochs with a batch size of 500. The Adam optimizer and a 0.001 learning rate were chosen during the training process. The training was conducted on an NVIDIA Tesla V100 DGXS GPU with 32 GB memory on a CentOS server with Intel Xeon E5-2698 CPU and 256 GB memory.

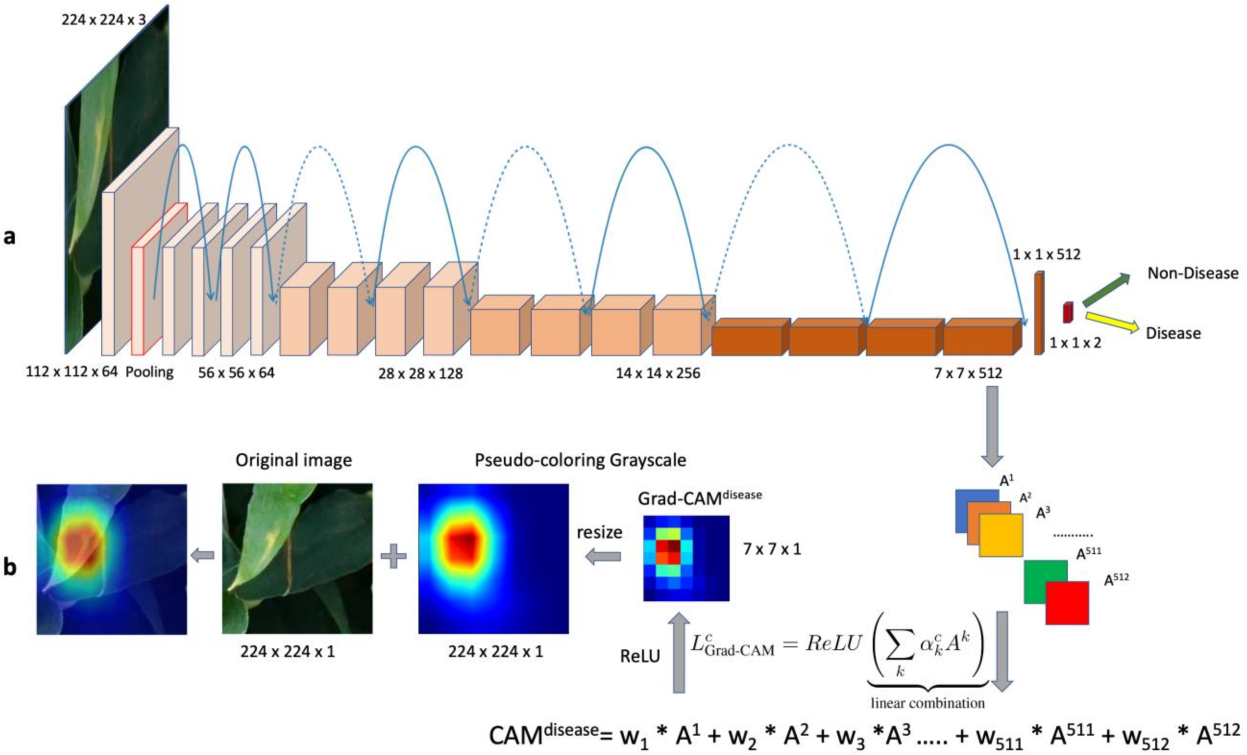


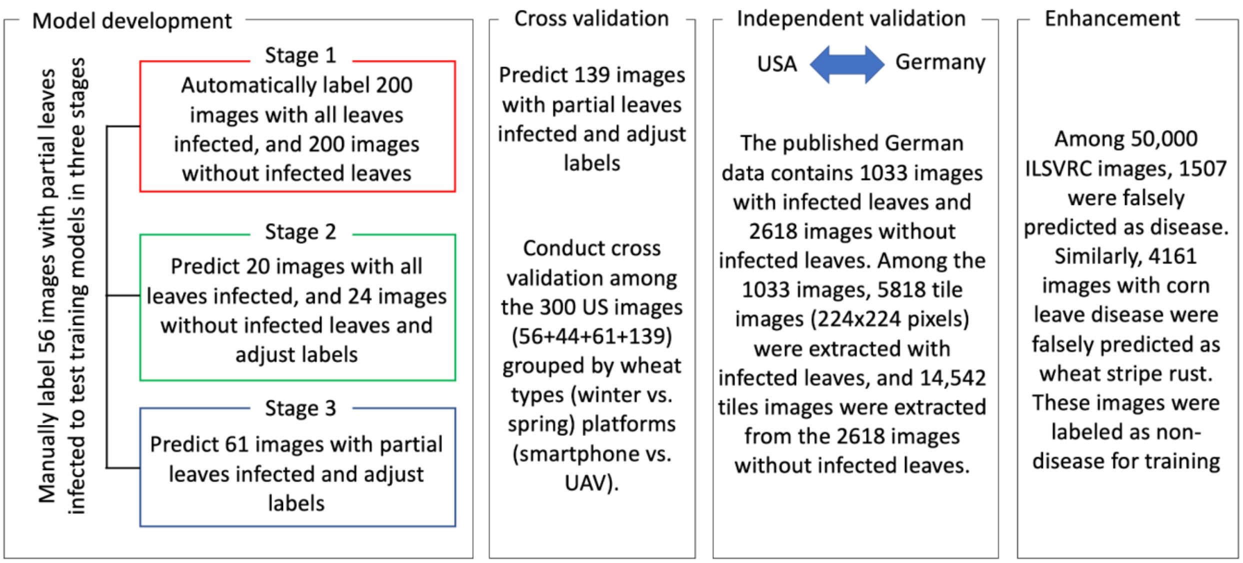
Figure 2. Architecture and visualization of the residual neural network-18. The pooling layer after the first convolutional layer was highlighted with red color. There are two types of shortcut connections in the architecture of the residual neural network-18 (a). The blue solid arrows show that the identical current layer was added with output of the next two convolutional layers, whereas

the dash arrows indicate a down-sampling of the current layer before the addition. A Grad-CAM (Selvaraju et al., 2020) was applied to visualize the last convolutional layer of ResNet-18, which highlights the regions that are critical for final prediction as disease or nondisease (b). A^k refers to the *k*th feature map from the last convolutional layer. α_k^c is identical to the weight (w_k) of A^k for the class *c*. The *ReLU* function will return zero for a negative input and keep a non-negative input as the same. A bilinear interpolation method was used to resize the image with smoothing.

A Gradient-weighted Class Activation Mapping (Grad-CAM) was implemented to visualize the ResNet-18 after the model was trained with all Pullman image data. The Grad-CAM can help to understand how the neural network makes a prediction for a specific class by highlighting the important region of an input image (Selvaraju et al., 2020). The Grad-CAM for the diseased class was calculated by the weighted addition of feature maps of the last convolutional layers. The weights for each feature map were calculated based on its gradient to the disease prediction, which was equal to the weights of the last fully connected layer. A *ReLU* function was applied to filter negative input (**Figure 2b**). A python package was used to visualize the Grad-CAM (https://github.com/jacobgil/pytorch-grad-cam).

Image labeling

Rooster software (https://github.com/12HuYang/Rooster) was used to label tile images into disease or non-disease classes by easily clicking it with a mouse. Rooster was developed with python and can split raw images into tiles (e.g., 224 × 224 pixels) by defining column and row numbers. A semi-automatic image labeling that combines machine- and human labeling was implemented in Rooster (**Figure 4g**). This semi-automated image labeling iteratively flows in a circle combining auto labeling, human adjustment, updating training set, and retraining classification model. The auto labels were predictions of the previously trained model, and human corrections were involved in adjusting incorrect classification. Images can be labeled in batches, and this stepwise system improved classification accuracy gradually (**Figure 3)**.



Model development

Manually label 56 images with partial leaves infected to test training models in three stages

Stage 1
Automatically label 200 images with all leaves infected, and 200 images without infected leaves

Stage 2
Predict 20 images with all leaves infected, and 24 images without infected leaves and adjust labels

Stage 3
Predict 61 images with partial leaves infected and adjust labels

Cross validation

Predict 139 images with partial leaves infected and adjust labels

Conduct cross validation among the 300 US images (56+44+61+139) grouped by wheat types (winter vs. spring) platforms (smartphone vs. UAV).

Independent validation

USA ⟷ Germany

The published German data contains 1033 images with infected leaves and 2618 images without infected leaves. Among the 1033 images, 5818 tile images (224x224 pixels) were extracted with infected leaves, and 14,542 tiles images were extracted from the 2618 images without infected leaves.

Enhancement

Among 50,000 ILSVRC images, 1507 were falsely predicted as disease. Similarly, 4161 images with corn leave disease were falsely predicted as wheat stripe rust. These images were labeled as non-disease for training

Figure 3. Overview of RustNet development. Image numbers in model development, cross validation, independent validation, and enhancement were summarized.

At Stage 1, tiles of 200 raw images without infected leaves and 200 images with almost all leaves infected were automatically labeled as non-disease (**Figure 4a**) and diseased (**Figure 4b**), which were indicated with a white or red top-and-left box, respectively. We used this initial training dataset to train the RustNet to distinguish disease and non-disease tiles. In Stage 2, tiles of 20 raw images with the majority of leaves infected (five still images from the UAV and 15 still images from the phone) were firstly predicted with the Stage 1 version of RustNet. Guided by the machine label, human input was involved to visually confirm or correct the label. With the updated label, these tiles were fed to RustNet to update the parameters. A similar iteration process was conducted for Stage 3. Tiles of 61 additional images with partial leaves infected (40 still images from the UAV and 21 still images from the phone) were predicted labels by the version Stage 2 of RustNet and then corrected by human input. With the updated label, these tiles were fed to retrain the RustNet.

For the testing set, a total of 56 labeled images were included in the testing set (49 still images from the UAV and seven still images from the phone). The RustNet versions that were trained with different training datasets in different labeling stages were compared using this testing set. With the RustNet version of Stage 3, the semi-automated labeling strategy was used to label 139 additional raw images. Finally, each category (still images from UAV and smartphone, video frame images from UAV and phone) from each farm (Spillman farm and PCFS) had 50 images (**Table 1**).

Images from the validation set of ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) (Russakovsky et al., 2015) were added to the training set as the non-disease class for negative control. This dataset has 50,000 images from 1000 classes which can expand domain knowledge of RustNet for the non-disease class. A corn leaf disease dataset (J & GOPAL, 2019; Singh et al., 2020) was also added to the training set for negative control. This corn leaf disease dataset consists of images from three corn leaf diseases (corn rust, gray leaf spot, and blight disease), which have a similar appearance to wheat stripe rust. Adding these two datasets for negative control can improve the specificity of RustNet when facing images from different domains.

Classification validation

Validation of classification of RustNet was conducted under three schemes, including image collection platforms, wheat types, and locations. These schemes included different ways to split data into training and testing data. The platform scheme and the wheat type scheme used the Pullman data. Under the platforms scheme, images from two of three platforms (images taken by phone, taken by UAV, or from videos) were used as training data and tested in the third platform (**Figure 3**). Under the wheat type scheme (spring wheat at the Spillman farm and winter wheat at the PCFS), images of one wheat type were used as training data and tested using images of the other wheat type (**Figure 3**). Under the location scheme, images were split by location (one-fold from Pullman, WA, USA, and another from Marquardt, Potsdam, Germany) (Schirrmann et al., 2021). Images from one location were used as training data and tested in the other location (**Figure 3**). Images of Marquardt were collected from different days after inoculation (DAI). Images of each DAI were tested with RustNet, which was trained with different parts of Pullman data.

Assessment of prediction accuracy

A prediction score of 0.5 or above is classified as case and control otherwise. The ground truth in testing datasets was used to derive the counts on true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Multiple criteria were calculated, including true positive rates (TPR), sensitivity, recall, specificity, false-positive rate (FPR), precision, false discovery rate (FDR), F1 score, and accuracy (ACC). Their formulas are summarized as follows.

$$\text{TPR} = \text{Sensitivity} = \text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{FPR} = \frac{FP}{TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{FDR} = \frac{FP}{TP+FP}$$

$$\text{F1 score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The cutoff of prediction scores was also varied from 0 to 1 to derive receiver operating characteristic (ROC) curves for TPR against FDR and FPR. The TPR of a random guess increases from 0 to 1 when the FPR increases from 0 to 1. Therefore, the area under the curve (AUC) is 0.5 under the null hypothesis. The curve of random guess is a straight line on the diagonal. However, the maximum of FDR depends on the ratio of control over the total (control + case). For example, when a testing data contains 100 cases and 100 controls, the maximum of FDR is 0.5. Therefore, the observed FDR should be divided by the observed maximum FDR to make the results comparable to the general situation with the maximum FDR of 1. The adjusted FDR has the property that the maximum FDR is 1. The minimum FDR of 0 was assumed for a TPR of 0. A random guess has a TPR of 0 until FDR is close to 1. Therefore, the AUC is 0 under the null hypothesis for ROC of TPR against FDR.

doi:10.20944/preprints202204.0177.v1

References

Chen, X. (2013). Review Article: High-Temperature Adult-Plant Resistance, Key for Sustainable Control of Stripe Rust. *American Journal of Plant Sciences*, *04*(03), 608–627. https://doi.org/10.4236/AJPS.2013.43080

Chen, X. (2020). Pathogens which threaten food security: Puccinia striiformis, the wheat stripe rust pathogen. *Food Security*, *12*(2), 239–251. https://doi.org/10.1007/S12571-020-01016-Z

Chen, X. M. (2005). Epidemiology and control of stripe rust [Puccinia striiformis f. sp. tritici] on wheat. *Canadian Journal of Plant Pathology*, *27*(3), 314–337. https://doi.org/10.1080/07060660509507230

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

J, A. P., & GOPAL, G. (2019). *Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network*. *1*. https://doi.org/10.17632/TYWBTSJRJV.1

Kang, Z., Li, X., Wan, A., Wang, M., & Chen, X. (2019). Differential sensitivity among Puccinia striiformis f. sp. tritici isolates to propiconazole and pyraclostrobin fungicides. *Canadian Journal of Plant Pathology*, *41*(3), 415–434. https://doi.org/10.1080/07060661.2019.1577301/SUPPL_FILE/TCJP_A_1577301_SM9836.DOCX

Liu, T., Wan, A., Liu, D., & Chen, X. (2017). Changes of races and virulence genes in Puccinia striiformis f. sp. tritici, the wheat stripe rust pathogen, in the United States from 1968 to 2009. *Plant Disease*, *101*(8), 1522–1532. https://doi.org/10.1094/PDIS-12-16-1786-RE/ASSET/IMAGES/LARGE/PDIS-12-16-1786-RE_F2.JPEG

Mi, Z., Zhang, X., Su, J., Han, D., & Su, B. (2020). Wheat Stripe Rust Grading by Deep Learning With Attention Mechanism and Images From Mobile Devices. *Frontiers in Plant Science*, *0*, 1386. https://doi.org/10.3389/FPLS.2020.558126

Oerke, E. C. (2020). Remote Sensing of Diseases. *Https://Doi.Org/10.1146/Annurev-Phyto-010820-012832*, *58*, 225–252. https://doi.org/10.1146/ANNUREV-PHYTO-010820-012832

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/S11263-015-0816-Y/FIGURES/16

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R., & Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. *Nature Reviews Genetics 2002 3:6*, *3*(6), 429–441. https://doi.org/10.1038/nrg817

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution 2019 3:3*, *3*(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

Schirrmann, M., Landwehr, N., Giebel, A., Garz, A., & Dammer, K.-H. (2021). Early Detection of Stripe Rust in Winter Wheat Using Deep Residual Neural Networks. *Frontiers in Plant Science*, *0*, 475. https://doi.org/10.3389/FPLS.2021.469689

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/S11263-019-01228-7/FIGURES/21

Shiferaw, B., Smale, M., Braun, H. J., Duveiller, E., Reynolds, M., & Muricho, G. (2013). Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Security*, *5*(3), 291–317. https://doi.org/10.1007/S12571-013-0263-Y/TABLES/9

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020). PlantDoc: A Dataset for Visual Plant Disease Detection. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. https://doi.org/10.1145/3371158

Su, J., Liu, C., Coombes, M., Hu, X., Wang, C., Xu, X., Li, Q., Guo, L., & Chen, W. H. (2018). Wheat yellow rust monitoring by learning from multispectral UAV aerial imagery. *Computers and Electronics in Agriculture*, *155*, 157–166. https://doi.org/10.1016/J.COMPAG.2018.10.017

Su, J., Yi, D., Su, B., Mi, Z., Liu, C., Hu, X., Xu, X., Guo, L., & Chen, W. H. (2021). Aerial Visual Perception in Smart Farming: Field Study of Wheat Yellow Rust Monitoring. *IEEE Transactions on Industrial Informatics*, *17*(3), 2242–2249. https://doi.org/10.1109/TII.2020.2979237

Wang, D. M., Wan, D. A., & Chen, D. X. (2022). Race Characterization of Puccinia striiformis f. sp. tritici in the United States from 2013 to 2017. *Https://Doi.Org/10.1094/PDIS-11-21-2499-RE*. https://doi.org/10.1094/PDIS-11-21-2499-RE

ZADOKS, J. C., CHANG, T. T., & KONZAK, C. F. (1974). A decimal code for the growth stages of cereals. *Weed Research*, *14*(6), 415–421. https://doi.org/10.1111/J.1365-3180.1974.TB01084.X

Zhang, C., & Kovacs, J. M. (2012). The application of small unmanned aerial systems for precision agriculture: A review. *Precision Agriculture*, *13*(6), 693–712. https://doi.org/10.1007/S11119-012-9274-5/FIGURES/3

Zhang, X., Han, L., Dong, Y., Shi, Y., Huang, W., Han, L., González-Moreno, P., Ma, H., Ye, H., & Sobeih, T. (2019). A Deep Learning-Based Approach for Automated Yellow Rust Disease Detection from High-Resolution Hyperspectral UAV Images. *Remote Sensing 2019, Vol. 11, Page 1554*, *11*(13), 1554. https://doi.org/10.3390/RS11131554