SCIENTIFIC REPORTS

OPEN

Early Detection of Tomato Spotted Wilt Virus by Hyperspectral Imaging and Outlier Removal Auxiliary Classifier Generative Adversarial Nets (OR-AC-GAN)

Dongyi Wang [ID][1], Robert Vinson[1], Maxwell Holmes[1], Gary Seibel[1], Avital Bechar[2], Shimon Nof[3] & Yang Tao[1]

Tomato spotted wilt virus is a wide-spread plant disease in the world. It can threaten thousands of plants with a persistent and propagative manner. Early disease detection is expected to be able to control the disease spread, to facilitate management practice, and further to guarantee accompanying economic benefits. Hyperspectral imaging, a powerful remote sensing tool, has been widely applied in different science fields, especially in plant science domain. Rich spectral information makes disease detection possible before visible disease symptoms showing up. In the paper, a new hyperspectral analysis proximal sensing method based on generative adversarial nets (GAN) is proposed, named as outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN). It is an all-in-one method, which integrates the tasks of plant segmentation, spectrum classification and image classification. The model focuses on image pixels, which can effectively visualize potential plant disease positions, and keep experts' attention on these diseased pixels. Meanwhile, this new model can improve the performances of classic spectrum band selection methods, including the maximum variance principle component analysis (MVPCA), fast density-peak-based clustering, and similarity-based unsupervised band selection. Selecting spectrum wavebands reasonably is an important preprocessing step in spectroscopy/hyperspectral analysis applications, which can reduce the computation time for potential in-field applications, affect the prediction results and make the hyperspectral analysis results explainable. In the experiment, the hyperspectral reflectance imaging system covers the spectral range from 395 nm to 1005 nm. The proposed model makes use of 83 bands to do the analysis. The plant level classification accuracy gets 96.25% before visible symptoms shows up. The pixel prediction false positive rate in healthy plants gets as low as 1.47%. Combining the OR-AC-GAN with three existing band selection algorithms, the performance of these band selection models can be significantly improved. Among them, MVPCA can leverage only 8 spectrum bands to get the same plant level classification accuracy as OR-AC-GAN, and the pixel prediction false positive rate in healthy plants is 1.57%, which is also comparable to OR-AC-GAN. This new model can be potentially transferred to other plant diseases detection applications. Its property to boost the performance of existing band selection methods can also accelerate the in-field applications of hyperspectral imaging technology.

Tomato spotted wilt virus (TSWV) is one of the common threats to more than 1,000 plant species from different botanical families[1]. It can cause a range of symptoms in a persistent and propagative manner, including sudden yellowing, mild mottling, mosaic and so on[2]. In practice, once the symptoms start developing, it is too late to head

[1]Bio-Imaging and Machine Vision Lab, Fischell Department of Bioengineering, University of Maryland, College Park, MD, 20742, USA. [2]The Institute of Agriculture Engineering, Agriculture Research Organization, Volcani Center, P.O.Box 6, Bet Dagen, 50250, Israel. [3]School of Industrial Engineering, Purdue University, 315N Grant Street, West Lafayette, IN, 47907-2023, USA. Correspondence and requests for materials should be addressed to Y.T. (email: ytao@umd.edu)

off an epidemic[3]. Since 2004, TSWV isolates have overcome the resistance gene Tsw in pepper[4], making it more difficult to manage. Bell pepper is a high-value specialty crop grown mostly in greenhouses for fresh markets. It is cultivated worldwide and used as a food ingredient, spice and ingredient in medicine. Therefore, early detection of TSWV is a crucial issue to ensure all the infected pepper plants eradicated as soon as possible[5].

Monitoring plant health and detecting pathogens effectively are important topics in precision agriculture research[6]. Early detection is meaningful to reduce disease spread and to facilitate management practice[7]. Molecular-level direct detection method can accurately evaluate the plant disease levels, but it's hard to be conducted in real-time field test scientific[8]. Comparatively, machine vision-based indirect detection method is more attractive in practice because of their non-invasive properties and their abilities to identify plant diseases through various parameters including color[9], morphological[10] and temperature changes[11]. Hyperspectral imaging (HSI) makes use of the plants' interaction with different electromagnetic spectra, and forms an image containing the intrinsic information of the leaf biochemical compounds and leaf anatomical structure[11]. Compared to RGB color imaging system, HSI including near infrared information makes early plant disease detection possible, and the subtle changes in spectral reflectance of plants could reflect early and invisible symptoms of the disease[12]. This technology has achieved great success in analyzing chemical component levels[13,14] and plant diseases[3,12,15].

However, problems which potentially constrain the wide applications of HSI in early stage plant disease detection remain to be solved. Firstly, most advanced HSI algorithms were developed and validated based on some public remote sensing dataset[16,17]. Furthermore, for early stage plant disease detection applications, it is very hard to get a well-prepared dataset, especially the pixel-level ground truth for invisible disease symptoms. Even experienced experts cannot label where the invisible disease symptoms are and define the pure invisible diseased pixel, which is important for some HSI analysis methods[18,19]. Most current research takes the plant as a whole, and the average plant spectrum is used for plant classification[20], whereas in practice, mean spectrum can hardly represent the whole plant. Different illumination conditions can make the spectrum of different plant locations vary significantly. For diseased plants in the early stage, the diseased spot may be small, and computing the average spectrum could wipe out the diseased symptoms. Meanwhile, because there is no prior knowledge of the spatial distribution of invisible symptoms, it is difficult to make use of spatial information to improve the HSI analysis performances though the strategy is very effective in many other HSI applications[19,21,22]. Therefore, determining exact spectrum characteristics is crucially important for hyperspectral image analysis and its applications for early plant disease detection.

Secondly, classifying healthy and diseased spectrums is not a trivial task because of the spectrum similarity and imaging system noise. Full spectra information may benefit the discrimination performance because there is no information loss, but meanwhile, it rapidly increases the complexity of modelling caused by information redundancy, especially with the required computation power and the analysis time in practice[23]. Decreasing the features number of the spectral signal is the common solution to the problem using techniques like projections[24], clustering[25–27], or autoencoders[28]. The extracted features can be sent into a discrimination model like linear discriminant analysis or support vector machine, to do classification[29,30]. However, some of the dimensional reduction processes sacrifice the original physical meaning of spectral signals due to some linear or non-linear transformations[31,32]. To preserve the physical information and make the model interpretable, selecting 'adequate' wavebands from the original hyperspectral space[3,33,34] is more attractive. For biological engineering applications, preserving the original band information is also very important because once the 'adequate' wavebands are determined, the expensive hyperspectral camera can be downgraded to more reliable and cost-effective multispectral camera which is more likely to be used in the field. There are many research on band selection algorithms, which can be divided into supervised and unsupervised methods[35–39]. Supervised band selection methods usually rely on some specific criterion functions and discrimination models[38,39]. However, for the early stage plant disease detection application, it is very difficult to quantify pixel-level classification performance of invisible diseased pixels, and thus unsupervised band selection is more promising in this specific domain. Most current unsupervised band selection models are easily affected by system noise, data crossovers and outliers[30], and how to improve their performance is still an open topic.

Consequently, a new hyperspectral analysis model, named as outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN) is proposed in this paper. It is a variant of Generative Adversarial Network (GAN)[40], a popular neural network architecture in deep learning domain. In recent years, the concept of deep learning has achieved great success in many areas[41,42]. In this paper, the proposed OR-AC-GAN is expected to meet the following specific objectives: (i) Without band selection, it can classify the pixels in hyperspectral images into backgrounds, diseased plant pixels and healthy plant pixels in single step, and thus the exact plant disease positions can be determined. (ii) From the plant level, it can detect diseased plants before specific symptoms show up. (iii) Compared to other models, it can effectively reduce the pixel prediction false positive rate in healthy plants without affecting the plant-level prediction result. (iv) It can generate some fake spectrum data following original data distribution. The generated data can be further used for some classic unsupervised band selection models and is expected to improve their performances. Overall, the novel method can detect the TSWV disease sensitively at an early stage before the symptom is visible making use of full hyperspectral information. It can also be potentially used in field for real-time applications because the generated data from the model can boost the performance of existing band selection algorithms and preserve the classification accuracy with limited bands.

Results

Outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN).

The idea of proposed OR-AC-GAN originates from a new and promising type of generative model named generative adversarial nets (GAN), as shown in Fig. 1[40]. It can learn the data distribution from scratch, and doesn't need any pre-knowledge and preliminary assumptions about dataset. Generally, there is a generator and a discriminator in the GAN model. The generator aims to create fake data as real as possible, and the discriminator targets for distinguishing
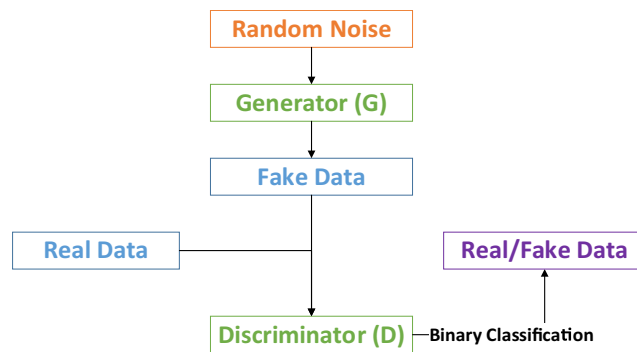
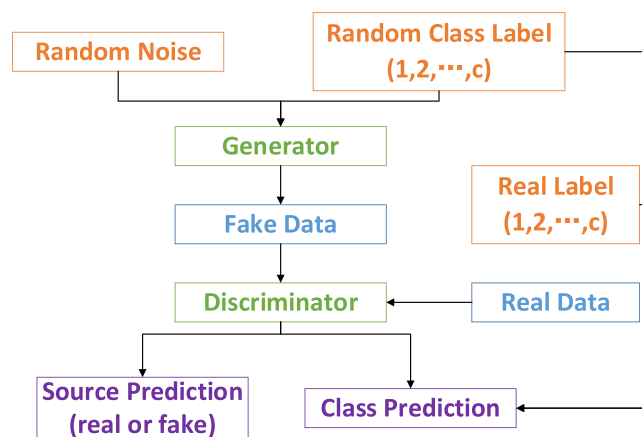**Figure 1.** Basic architecture of GAN model.

Figure 2. The architecture of AC-GAN model.

the fake date g from the real data x. Once the GAN model is well-trained, the generator is expected to describe real data distribution, which can be used for augmenting the dataset.

Auxiliary classifier GANs (AC-GAN) is a variant of GAN network[43], which combines a c classification task into the GAN model, as shown in Fig. 2. It can effectively augment the classification dataset during the network training procedure. In addition, research shows the additional task can stablize the GAN training process[44,45].

In practice, the Achilles' heel of AC-GAN model is if the real distribution of two classes is very closed, the data augmentation by generator can completely ruin the classification ability of model. The reason for that is, in AC-GAN model, even if the binary discriminator determines the data as fake, the classifier still needs to allocate the data into a particular class. It is originally designed for increasing the dataset, but it also strengthens the side-effect of data outliers and crossovers. This problem becomes overwhelming in our application due to the trival spectrum differences among healthy and diseased pixels. The augmentated data of the two different classes could confuse the *D*, and further affect the generation ability of *G*.

The proposed OR-AC-GAN made a subtle change of AC-GAN, but the inherent idea of the two models are completely different. As shown in Fig. 3, the art of OR-AC-GAN is that an additional label *c + 1* is allocated when training the *D*, and all fake data is classified into the additional class. It means even if the fake data is closed enough to the real data, it will still be classified into the additional type. This design can obviously improve the classification criterion, rule out the data outliers, and the generated data from OR-AC-GAN can also focus on the intrinsic features of data in different classes. In the test phase, the *D* can classify the image spectrum pixels as background, healthy or TSWV, which can be used for locating the diseased positions.

Figure 4 shows some typical spectrums of healthy, TSWV and background pixels in real dataset. After the OR-AC-GAN model is well-trained, the generated spectrums which are shown in Fig. 5 can capture the intrinsic features of real spectrums in different classes. The generated spectrums improve the classification results of OR-AC-GAN and the performance of classic band selections models. The related experiment results will be shown in following sections. The well-trained discriminator can segment the hyperspectral images as described in Fig. 3. The typcial visualized classificaiton results are shown in Fig. 6, where green indicates healthy and red indicates the possible TSWV infection.

Evaluation of pixel and plant level classification results. As shown in Fig. 6(j), there are some predicted diseased pixels in a healthy plants. To futher quantify the classification performance of model, two pixel-level and two plant-level metrics are defined here.
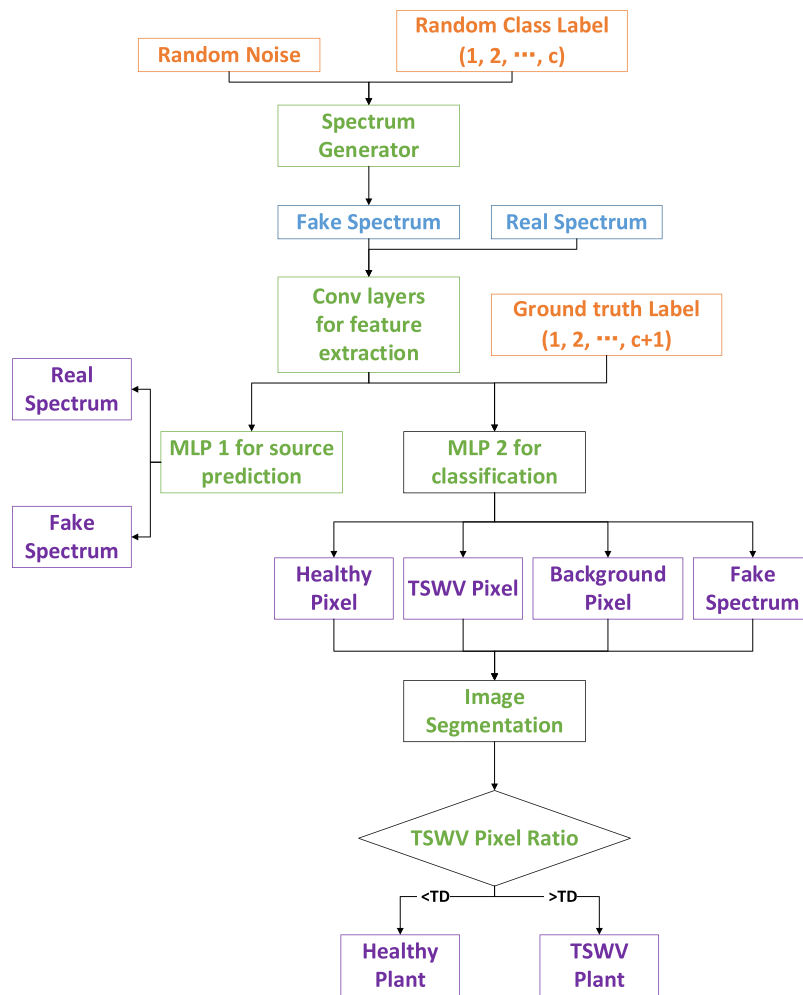
Figure 3. The system architecture of proposed OR-AC-GAN model and its application for early TSWV detection. MLP 1 and MLP 2 are two multi-layer perceptrons. TSWV (diseased) pixel ratio is defined as the ratio of the number of predicted TSWV pixels to the sum of predicted TSWV and healthy pixels. *TD* means a threshold to determine the plant is diseased or not based on the TSWV pixel ratio.
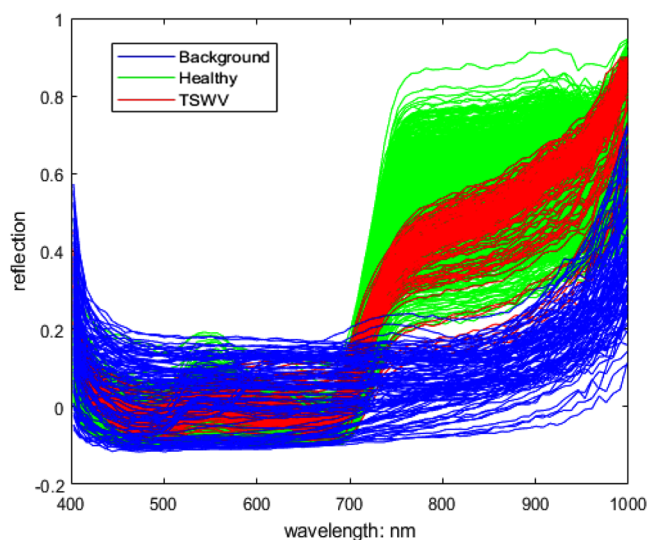

Figure 4. The typical real spectrums of different classes. Blue: background pixel. Green: healthy pixels. Red: TSWV pixels.
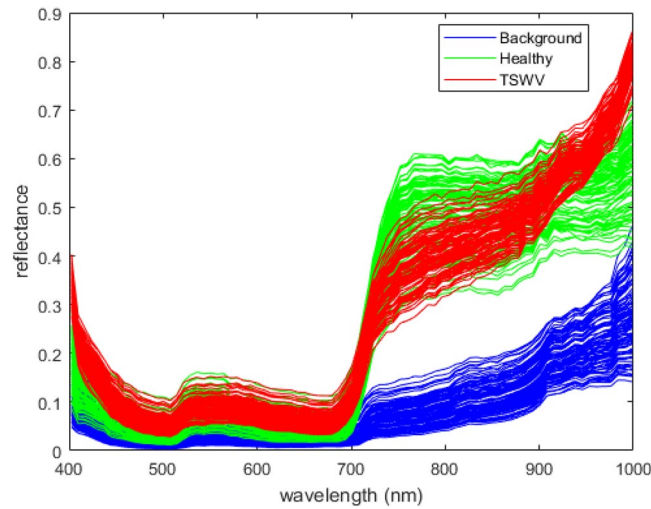
Figure 5. The typical generated spectrums of different classes. Blue: fake background pixels. Green: fake healthy pixels. Red: fake TSWV pixels.
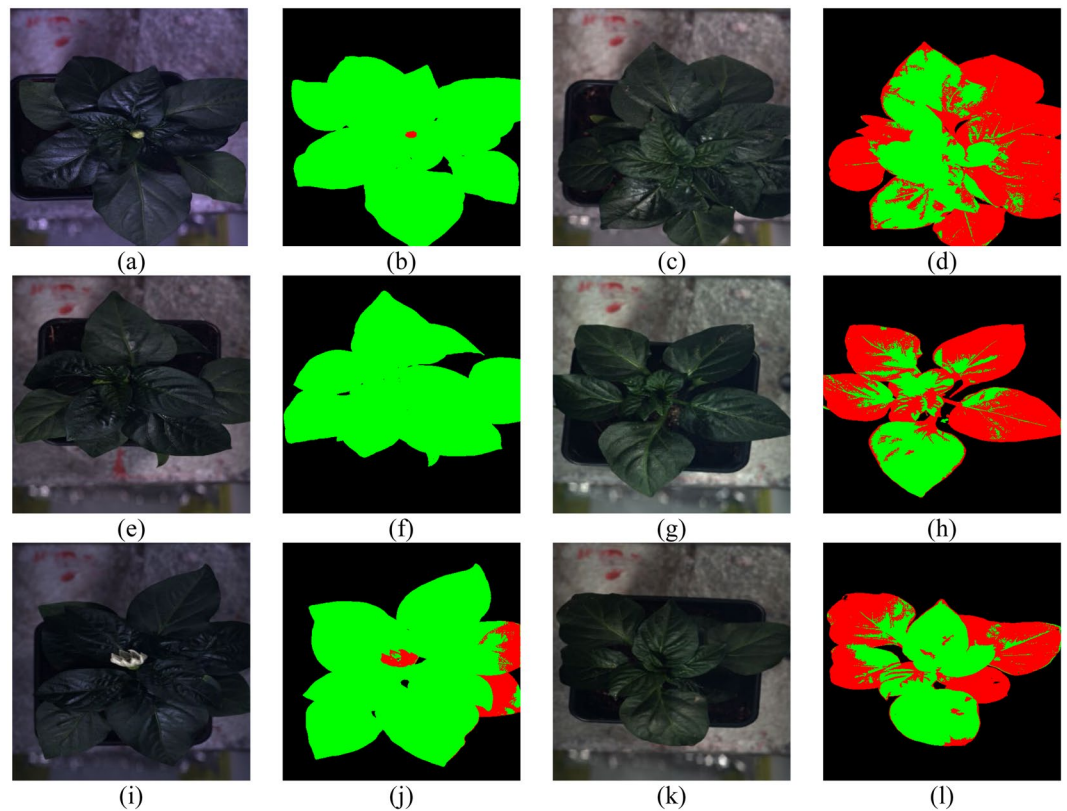

Figure 6. Typical results for test plants from OR-AC-GAN. Healthy pixels are labelled as green, and TSWV pixels are labelled as red. Left: (**a**,**b**), (**e**,**f**), and (**i**,**j**) are three healthy plants and their corresponding classification results. Right: (**c**,**d**), (**g**,**h**), and (**k**,**l**) are TSWV infected plants and their corresponding classification results.

The first pixel-level metric is to evaluate the plant segmentation performance. The TSWV and healthy pixels are called by a joint name plant pixels. The ground truth plant pixels are labelled manually. The expression of the metric, Acc_{pixel}, is defined in Equation 1.

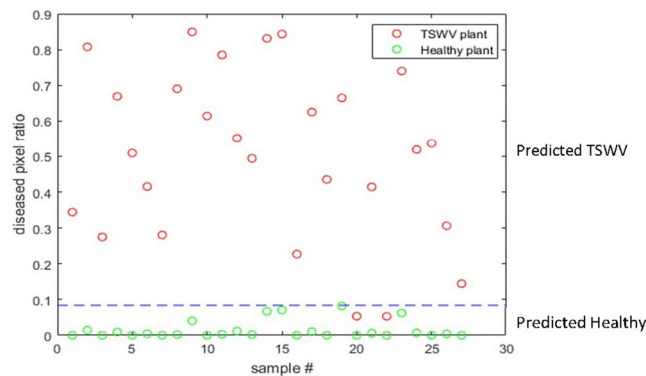$$\text{Acc}_{pixel} = \frac{PNC_{ppixel} + PNC_{bpixel}}{PN_{total}}$$

(1)

Figure 7. The classificaiton result of an independent test dataset, where the predicted healthy and diseased plants can be separated by the threshold *TD* determined by the algorithm in Fig. 3 based on the diseased pixel ratio of a plant.

where PNC_{ppixel} is the number of plant pixels which are predicted correctly (TSWV and healthy pixels may mix up). PNC_{bpixel} is the number of background pixels which are predicted correctly. PN_{total} is the number of pixels in a hyperspectral image. The average Acc_{pixel} value is 98.03% for 54 plants in test dataset.

The second pixel-level index is to evaluate how well the model can distinguish the healthy pixels and TSWV pixels. It is defined as the false positive rate of TSWV pixels in the healthy plants ($FP_{Tpixel,Healthy}$), shown in Equation 2.

$$FP_{Tpixel,Healthy} = \frac{PN_{Tpixel,Healthy}}{PN_{Tpixel,Healthy} + PN_{hpixel,Healthy}} \tag{2}$$

where $PN_{Tpixel,Healthy}$ is the number of predicted TSWV pixels in healthy plants, $PN_{hpixel,Healthy}$ is the number of predicted healthy pixels in healthy plants. This metric is only defined in healthy plants because pixel-level ground truth is not available in the TSWV plant, but all the plant pixels in the healthy plants should be classified as healthy. For the 27 healthy plants in test dataset, the average $FP_{Tpixel,Healthy}$ value is 1.47%, with standard derivation of 2.53%. The worst (largest) value in the test dataset is 8.21%.

The plant-level metric includes the specificity and sensitivity value based on the plant-level classification results. The definitions are shown in Equation 3.

$$\begin{cases} Sensitivity_{plant} = \dfrac{TP_{plant}}{TP_{plant} + FN_{plant}} \\ Specificity_{plant} = \dfrac{TN_{plant}}{TN_{plant} + FP_{plant}} \end{cases} \tag{3}$$

where TP_{plant} is the true positive value. TN_{plant} is the true negative value. FP_{plant} is the false positive value, and FN_{plant} is the false negative value. All these metrics are defined in plant-level.

Figure 7 shows the result of an independent test dataset of 54 plants by the algorithm in Fig. 3. In Fig. 7, the red points represent diseased plants, and the green points represent healthy plants. The blue line represents the disease threshold *TD* defined in Fig. 3. The plants with diseased pixel ration larger than *TD* are defined as TSWV plants, otherwise they are predicted as healthy plants. The *TD* value determination takes both sensitivity and specifity into consideration and aims to maximize the quadratic sum of sensitivity and specificity. In the experiment, the optimized *TD* value is 0.084 and the corresponding sensitivity and specificity value are 92.59% and 100%, respectively.

Comparison with one dimensional convolutional neural network (1D-CNN) and AC-GAN. One dimensional CNN has been a powerful and promising method in spectrum analysis domain in recent years[46,47]. However, in the plant diseases detection applications, imbalance number diseased and healthy pixels can affect the classification results. As mentioned earlier, AC-GAN can augment the classification dataset, but it can also amplify the outliers in the dataset. The proposed OR-AC-GAN aims to solve the problems in the two current models. With the same network configurations, the comparison results of three models are shown in Table 1.

For Acc_{pixel}, there are no significant differences among three models, and all of them get >98% classification accuracy. It means the three models can successfully distinguish plant pixels and background pixels. For $FP_{Tpixel,healthy}$, healthier pixels are regarded as diseased pixels in AC-GAN than 1D-CNN proving the statement that AC-GAN can intensify the side-effects of outliers, and decrease the discrimination ability of model. On the contrary, the OR-AC-GAN can weaken the side-effects and augment the dataset online. Figure 8 shows the segmentation result of a typical healthy plant based on the three models, which supports the data shown in Table 1.

	Acc_{pixel}	$FP_{Tpixel,Healthy}$ (average)	$FP_{Tpixel,Healthy}$ (standard derivation)	$FP_{Tpixel,Healthy}$ (worst)	$Sensitivity_{plant}$	$Specificity_{plant}$
1D-CNN	98.02%	5.95%	14.79%	34.75%	**92.59%**	92.59%
AC-GAN	98.00%	10.41%	15.32%	57.59%	88.89%	**100%**
OR-AC-GAN	**98.03%**	**1.47%**	**2.53%**	**8.21%**	**92.59%**	**100%**

Table 1. Statistic comparison results of different network architectures.

Band Number	1	2	3	4	5	6	7	8
MVPCA	0.430	0.336	0.214	0.143	0.139	0.115	0.112	0.123
OR-AC-GAN + MVPCA	0.542	0.151	0.146	0.101	0.053	0.066	0.017	**0.016**
FDPC	0.175	0.158	0.148	0.106	0.089	0.075	0.083	0.088
OR-AC-GAN + FDPC	0.091	0.091	0.090	0.076	0.077	0.071	0.070	**0.069**
SUBS	N/A	0.204	0.081	0.084	0.077	0.078	0.078	0.080
OR-AC-GAN + SUBS	N/A	0.245	0.073	0.074	0.074	0.070	**0.063**	0.071

Table 2. The pixel-level false positive rate ($FP_{Tpixel,Healthy}$) of MVPCA, FDPC, SUBS, OR-AC-GAN + MVPCA, OR-AC-GAN + FDPC and OR-AC-GAN + SUBS.
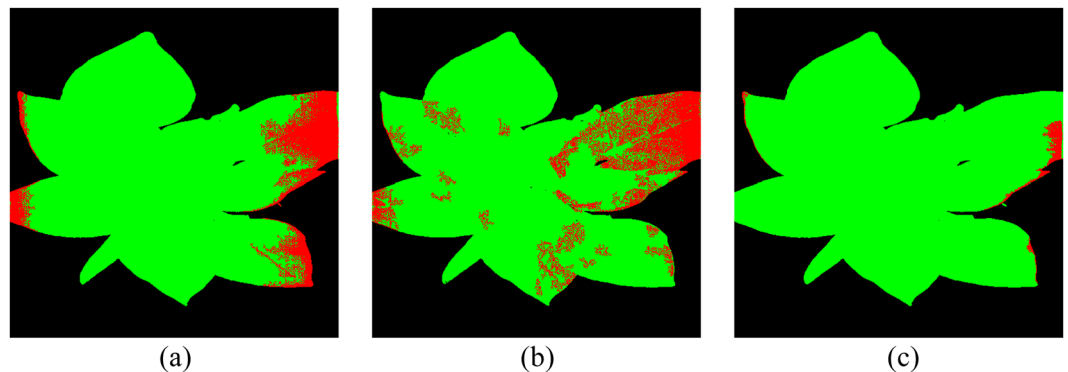

Figure 8. Comparison of segmentation results of a typical healthy plant. (**a**) direct CNN model. (**b**) AC-GAN model. (**c**) the proposed OR-AC-GAN model.

Improve the performances of classic band selection algorithms.   Three classic unsupervised band selection algorithms are tested here, including maximum variance principle component analysis (MVPCA)[37], fast density peak-based clustering (FDPC)[36] and similarity-based unsupervised band selection (SUBS)[35] models. The experiment wants to prove OR-AC-GAN can improve the performance of the band selection models. All these models are firstly applied based on the real spectrum dataset. Then the dataset is doubled by adding the fake spectrums from OR-AC-GAN, and all the band selection algorithms are conducted again based on the augmented dataset. Once the spectrum bands are determined, the pixels will be classified as background, TSWV or healthy pixels via the k-nearest neighbor algorithm (KNN)[48]. KNN is a classic machine learning algorithm which has been widely applied in scientific and engineering domain[49,50].

Table 2 shows how the number of selected bands affect the average value of $FP_{Tpixel,healthy}$ in different band selection models. Generally, the average $FP_{Tpixel,healthy}$ value decreases with more bands are selected for classification. Compared to applying the band selection on the original dataset, the band selection based on both original and fake dataset can effectively lower the $FP_{Tpixel,healthy}$ value. This trend usually shows up with more spectrum bands are selected, because the data distributions from OR-AC-GAN are more concentrated and KNN can't get effective information from one or two bands of the refined data. Among the three classic band selection models, MCPCA gets smallest $FP_{Tpixel,healthy}$ value (1.57%), which is comparable to the OR-AC-GAN result (1.47%). Eight bands are selected by MVPCA, which are 522 nm, 572 nm, 629 nm, 643 nm, 694 nm, 797 nm, 804 nm and 908 nm. The KNN model considers 4 nearest neighbors in this case. The plant segmentation result of typical plants are shown in Fig. 9. The Acc_{pixel} value is 97.38%, also comparable to the 98.03% from OR-AC-GAN. For the plant-level, the $Sensitivity_{plant}$ and $Specificity_{plant}$ values are same as the OR-AC-GAN results, which are 92.59% and 100%, respectively.

From the view of timelines.   The hyperspectral images are taken 5 day after inoculate (d.a.i), 7 d.a.i and 13 d.a.i separately. In this section, the analysis for OR-AC-GAN will be conducted from the view of timeline. The changes of the TSWV diseased pixel ratio to the plants (excluding the plants used for training) pixel (*TPR*) with time are shown in Fig. 10. The *TPR* for healthy plants is same as the definition of $FP_{Tpixel,healthy}$.
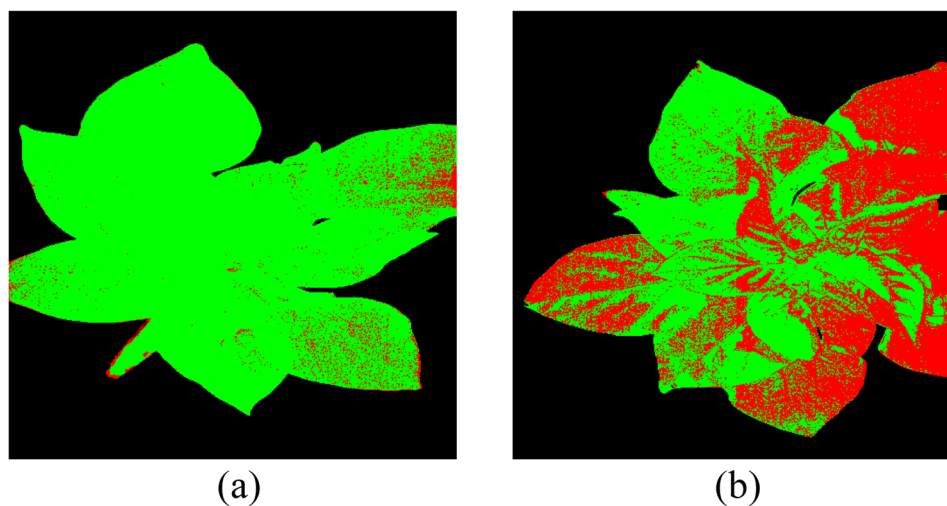
Figure 9. Typical visible prediction results from OR-AC-GAN + MVPCA + KNN. (**a**) A healthy plant (**b**) a TSWV plant.
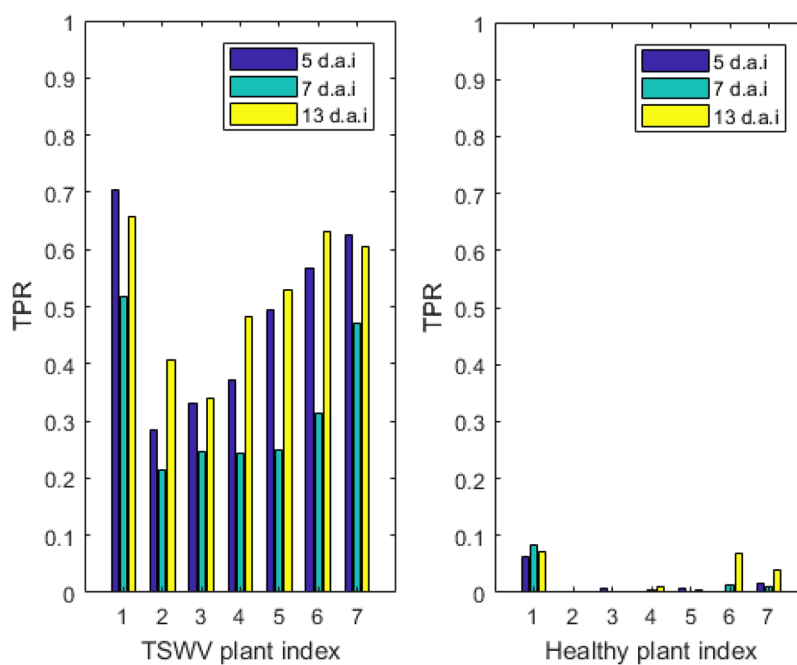

Figure 10. The TSWV diseased pixel ratio to the plant pixels (TPR value) for TSWV and healthy plants in different d.a.i. The three TSWV plants and three healthy plants in the training dataset are not included in the figure.

In the figure, the differences among healthy and TSWV plants are obvious and have been discussed in previous sections. The *TPR* values of both healthy and TSWV plants maintain the consistency in different d.a.i. For TSWV plants, the images from 13 d.a.i show relatively high *TPR* compared to the images from 5 and 7 d.a.i. The *TPR* relationships of 5 and 7 d.a.i images are upside-down compared to the expectation. There are couple of potential reasons to explain the problem. Firstly, there are some new leaves showing up in 7 d.a.i, which could block the old leaves and directly affect the *TPR* value. Secondly, other research[20] shows, for the particular plant disease, there are some crossovers of hyperspectral spectrums in different d.a.i, and TSWV could have similar phenotypes in spectrum domain. Thirdly, in network training process, the time information is not included in the pixel-level ground truth, because during the data labelling process, human experience can only determine the plant status according to visible disease symptoms. Lastly, illumination condition could affect the classification result. Further research will continue to explain the phenomenon deeply. Nevertheless, the OR-AC-GAN model has revealed the significant comparative difference between the diseased and healthy as early as 5 d.a.i.
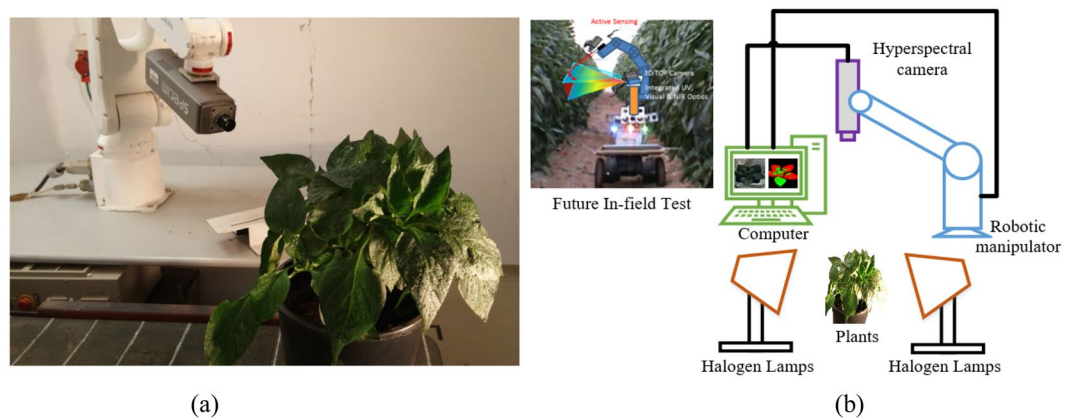
Figure 11. (**a**) The real experiment imaging station. (**b**) The schematic diagram of the image stataion. The system is expected to be mounted in the agriculture robotic for in-filed plant disease detection.

Discussion

Targeting for the early stage plant disease detection applications, a new hyperspectral analysis model, OR-AC-GAN, is proposed. In this report, a wide-spread plant disease TSWV is used for validating the model. The pixel-level classification false positive rate in healthy plants can achieve as low as 1.47%. The plant-level classification sensitivity and specificity can get 92.59% and 100%. The average classification accuracy is 96.25%.

Compared to existing research[12,20,51,52], the proposed model have serveral advantages. Firstly, traditional analysis models need firstly determine region of interests in images and extract 'reasonable' image features, including spectrum bands and spatial features. All these information was sent into a classifier to conduct pixel-level or plant-level classification. For different environment and experiment objects, the analytical strategies could vary a lot. On the contrary, OR-AC-GAN is a relatively fixed all-in-one model, which successfully integrates the task of image segmentation, feature extraction and classification.

Secondly, compared to applying CNN model directly, OR-AC-GAN successfully augment the dataset online, especially the data of diseased pixels. Compared to the AC-GAN model, it can automatically remove the outliers in the dataset, and avoid the decreasing discrimination ability accompanying with the data augmentation.

Thirdly, the well-trained OR-AC-GAN model is able to generate fake spectrums from random noises. Combining with the real spectrums, the fake spectrums improve the robustness and performance of some classic band selection methods. The trend becomes remarkable with the increasing of the number of selected spectrum bands. In the experiment, three band selection methods are tested, including MVPCA, FDPC and SUBS. Among them, MVPCA shows the best performance. It uses eight wavelengths to get comparable results as OR-AC-GAN which utilizes 83 wavebands. This experiment is meaningful from both the scientific and engineering view. Pre-determination of the spectrum bands can not only cut the cost of analytical system but also make the hyperspectral model explainable from the scientific view. For example, the eight bands selected by MVPCA are highly related to the photosynthetic capacity and red inflection point according to other research[53].

Our further research will focus on the view of timeline, and observe how hyperspectral images changes with time before visible symptoms in plants show up. Currently, the experiment only proves TSWV is distinguishable as early as 5 d.a.i. Meanwhile, different diseases will be tested to prove the robustness of the model. To further improve the early stage disease detection efficiency, other information like leaf temperature, chlorophyll content is expected to be integrated into the proposed model.

Methods

Image dataset construction. Plants of sweet pepper (Hazera Genetics) were obtained from a commercial nursery (Hishtil, Ashkelon, Israel) 40–50 days after seeding and were transplanted into 20 pots containing soil and potting medium and were fertigated proportionly with drippers 2–3 times per day with 5:3:8 NPK fertilizer (nitrogen (N), phosphorus (P) and potassium (K)), allowing for 25–50% drainage. Ten healthy plants (control), 10 plants infected with TSWV. Infecting the diseases were conducted and controlled by a plant pathologist. Images of the top part of all plants were acquired at a laboratory with hyperspectral camera (400–1000 nm, V10E Specim ImSpector) mounted on a Motorman 5L robotic manipulator as shown in Fig. 11(a). Two halogen lamps were placed within 0.5 m from the examined plant with a vertical orientation of 45 degrees as light sources. The schematic of the imaging system is shown in Fig. 11(b). Hyperspectral images were taken 5 d.a.i. (days after infection), 7 d.a.i. and 13 d.a.i, and total 60 images are in the dataset. All the images are calibrated based on the white and dark reference image[20].

Network training procedure. The dataset consists 60 hyperspectral images, 30 of which are healthy plants images, and the left 30 are TSWV plants images plants. To train the OR-AC-GAN model, a pixel-level spectrum training dataset needs to be prepared. Because most of TSWV plants haven't shown visible symptoms in RGB images, the three TSWV plant images in the training dataset need to be well-chosen and the diseased spots in the selected plants should be visible. In the experiment, three training TSWV images are selected from the data of 13 d.a.i. The diseased pixels in the plants are manually labelled and added into the pixel-level training dataset.

On the contrary, there is no specific requirement for selecting training healthy plants. In the experiment, three healthy plants are randomly selected from the data of 5 d.a.i., 7 d.a.i., 13 d.a.i, respectively. To rule out the spectrum differences resulting from different illumination conditions, the pixels in both the common illumination and the shadow are expected to be included in the training dataset. Background pixels in the pixel-level training dataset are selected from three TSWV images and three healthy images mentioned above. In total, there are 103,769 background pixels, 105,561 healthy pixels and 2,071 TSWV pixels in the pixel-level training dataset. The typical spectrums of the three type of pixels in shown Fig. 4. The remaining 54 images are used for algorithm test.

The well-prepared pixel-level dataset is used for training OR-AC-GAN model. The inputs of generator in OR-AC-GAN model are 40 random numbers ranging from 0 to 1 which follow the uniform distribution. The random noise passes through a series of fully connected layers, convolutional layers and non-linear operations to acquire a fake spectrum vector with 83 bands. The detailed structure of generator is shown in Fig. 12.

Both the generated spectrum and real spectrum are fed into the discriminator in OR-AC-GAN model. The discrimator is composed of several one-dimensional convolutional layers to extract data features. These features are sent into multiple layer perceptron (MLP) in parallel for two different tasks, source prediction and classification. The MLP for source prediction is equipped with only one fully connected layer and the MLP for classification has two fully connected layers in series. This design is to separate the two tasks based on the degree of difficulties. The detailed structure of discriminator is shown in Fig. 13.

Define the log-likelihood of source prediction as L_S, the log likelihood of classification in D training process as L_{C1}, and the log-likelihood of classification in G training process as L_{C2}. The definitions of L_S, L_{C1} and L_{C2} are shown in Equation 4. When D is trained, the network's target is to maximize $L_S + L_{C1}$. It aims to determine the fake data as fake and real data as real. Meanwhile, it needs to classify the data into correct classes. Real data is classified according to the labelled ground truth, and the fake data is classified into the additional class $c + 1$. When training G, the network's target is to maximize $L_{C2} - L_S$. The role of L_{C2} is to classify the real data corresponding to the labelled ground truth, same as L_{C1}. However, it also needs to classifty the fake data according to the random class labels fed into the generator.

$$
\begin{cases}
L_S = E_{x \sim p_x}[\log P(S = real|x)] + E_{z \sim p_z}[\log P(S = fake|G(z))] \\
L_{C1} = E_{x \sim p_x}[\log P(C = c_{real}|x)] + E_{z \sim p_z}[\log P(C = c + 1|G(z))] \\
L_{C2} = E_{x \sim p_x}[\log P(C = c_{real}|x)] + E_{z \sim p_z}[\log P(C = c_{fake}|G(z))]
\end{cases}
\tag{4}
$$

The training process is based on the principle of back propagation and gradient descent algorithm[48]. The learning rate is updated based on Adam method[54]. Due to the imbalance of the dataset, the class weights are adjusted based on the number of training pixels in different class. After 50-epoch training, the output of generator is visualized as in Fig. 5. The deep learning codes are implemented by the Python keras library[55] with Nvidia GeForce GTX Titan Xp GPU.

Band Selection. OR-AC-GAN model utilizes 83 spectrum bands information to do the pixel-level classification. However, in practice, if some spectrum bands can be selected in advance, the analysis model will be dramatically simplified, and the hyperspectral spectral imaging system can also be degraded into multispectral imaging system to cut the cost.

There are classic hyperspectral bands selection algorithms posed in remote sensing field. The main target of band selection criterial is to reduce the data redundancy and to select representative examplars. However, these band selection methods are highly relied on the data quality. Similar spectrums and data outliers can affect the performances of band selection algorithms. As mentioned earlier, a well-trained OC-AC-GAN model can remove data outliers online, and it is expected to improve the performance of three classic band selection models, including MVPCA[37], FDPC[36] and SUBS[35]. The descriptions of the three models are listed below. Assuming there are N one-dimensional spectrum data with l bands. The s_{ij} means the jth band of ith data sample.

Maximum variance principle component analysis (MVPCA). MVPCA[37] is a joint band prioritization and band decorrelation approach. It ranks the bands by a criterion that comprises the importance of an individual band and its correlation with other bands[36].

Assume $\Sigma = 1/N\sum_{i=1}^{N}(s_i - m)(s_i - m)^T$ is the covariance matrix of spectrum data, where m is the sample mean vector. The importances of each band is defined in Equation 5.

$$
Weight_k = \sum_{i=1}^{l} r_{ik}^2 \quad k = 1, 2, …, l
\tag{5}
$$

where $r_{ik} = \sqrt{\lambda_i} v_{ik}$, λ_i is the ith eigenvalues of Σ and v_{ik} is the kth value of ith eigenvector of Σ.

Fast density peak-based clustering (FDPC). FDPC[36] utilizes the two reasonable assumptions to create a metric to describe the importances of different spectrum bands. A good examplar should has high local density and relatively large distance from points of higher density[56]. It regards all the pixel values of the ith band, as a new data, noted as $s_{.i}$. The local density of the ith band, and its relatively distance to the higher density are described by ld_i and dh_i, as shown in Equation 6.

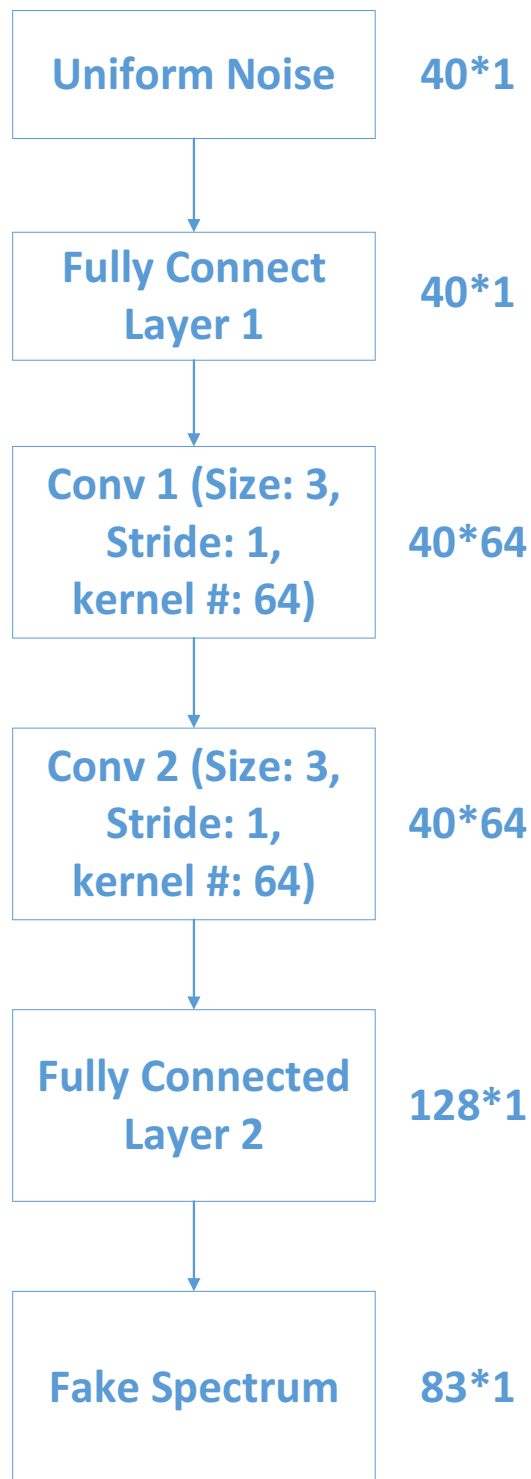| Uniform Noise | 40*1 |
Fully Connect Layer 1	40*1
Conv 1 (Size: 3, Stride: 1, kernel #: 64)	40*64
Conv 2 (Size: 3, Stride: 1, kernel #: 64)	40*64
Fully Connected Layer 2	128*1
Fake Spectrum	83*1

Figure 12. The detailed generator structure in proposed OR-AC-GAN model.

$$\begin{cases} ld_i = \sum_{j=1}^{l} \chi(d_{ij} - d_c) \\ dh_i = \min_{j:ld_j > ld_i} (d_{ij}) \end{cases}$$

(6)

where d_{ij} is the Euclidean distance (2-norm operator) between the ith spectrum band $s_{:i}$ and jth spectrum band $s_{:j}$, defined as $d_{ij} = \left\| s_{:i} - s_{:j} \right\|_2$. χ is a function of the difference value between d_{ij} and the cutoff distance d_c. If its input
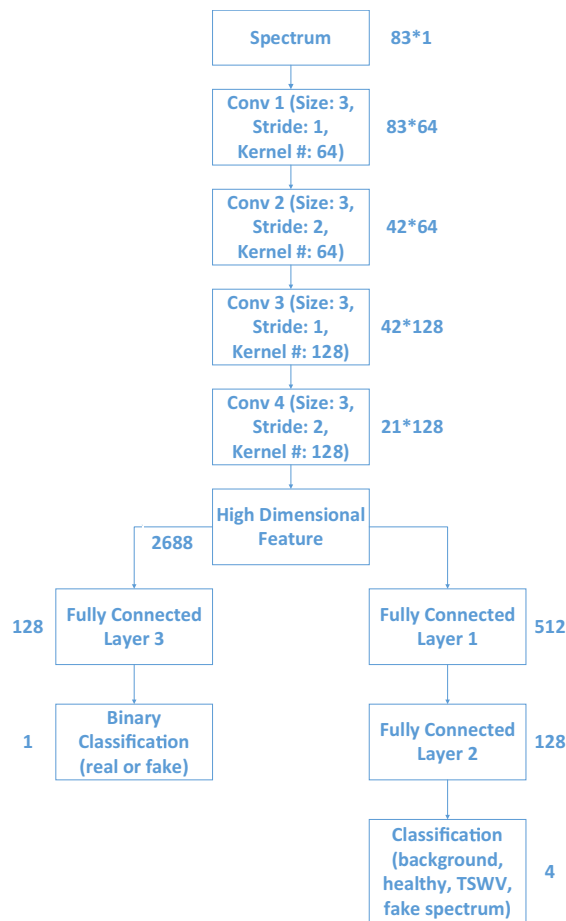
Figure 13. The detailed discriminator structure in proposed OR-AC-GAN model.

is negative, the function value is 1. Otherwise, the function value is 0. In FDPC, the importances of each band are defined as Equation 7.

$$Weight_k = \sum_{k=1}^{l} ld_k * dh_k \quad k = 1, 2, \ldots, l \tag{7}$$

Similarity-based unsupervised band selection (SUBS). SUBS[35] is a sequential forward search algorithm to achieve band selection. The algorithm starts from two initial bands determined by the maximum projection algorithm. Then SUBS assumes all other bands can be estimated linearly by the existing bands. The new selected band should lead to the largest prediction error. The process continuous until the number of selected bands meet the target.

Data Availability
The data that support the findings of this study are available from the corresponding author upon reasonable request.

References
1. Roselló, S., Díez, M. J. & Nuez, F. Viral diseases causing the greatest economic losses to the tomato crop. I. The Tomato spotted wilt virus—a review. *Scientia Horticulturae* **67**, 117–150 (1996).
2. Fereres, A. & Raccah, B. Plant virus transmission by insects. *eLS* (2015).
3. Krezhova, D., Petrov, N., & Maneva, S. Hyperspectral remote sensing applications for monitoring and stress detection in cultural plants: viral infections in tobacco plants. In Remote Sensing for Agriculture, Ecosystems, and Hydrology XIV (Vol. 8531, p. 85311H). International Society for Optics and Photonics (2012, October).
4. Margaria, P., Ciuffo, M. & Turina, M. Resistance breaking strain of Tomato spotted wilt virus (Tospovirus; Bunyaviridae) on resistant pepper cultivars in Almeria, Spain. *Plant Pathology* **53**, 795–795 (2004).
5. Avila, Y. *et al.* Evaluation of Frankliniella bispinosa (Thysanoptera: Thripidae) as a vector of the Tomato spotted wilt virus in pepper. *Florida Entomologist* **89**, 204–207 (2006).
6. Zhang, N., Wang, M. & Wang, N. Precision agriculture—a worldwide overview. *Computers and electronics in agriculture* **36**, 113–132 (2002).
7. Martinelli, F. *et al.* Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development* **35**, 1–25 (2015).
8. Fang, Y. & Ramasamy, R. P. Current and prospective methods for plant disease detection. *Biosensors* **5**, 537–561 (2015).

9. Barbedo, J. G. A., Koenigkan, L. V. & Santos, T. T. Identifying multiple plant diseases using digital image processing. *Biosystems Engineering* **147**, 104–116 (2016).
10. Camargo, A. & Smith, J. An image-processing based algorithm to automatically identify plant disease visual symptoms. *Biosystems engineering* **102**, 9–21 (2009).
11. Mahlein, A.-K., Oerke, E.-C., Steiner, U. & Dehne, H.-W. Recent advances in sensing plant diseases for precision crop protection. *European Journal of Plant Pathology* **133**, 197–209 (2012).
12. Rumpf, T. *et al.* Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture* **74**, 91–99 (2010).
13. Li, F. *et al.* Improving estimation of summer maize nitrogen status with red edge-based spectral vegetation indices. *Field Crops Research* **157**, 111–123 (2014).
14. Zhai, Y. *et al.* Estimation of nitrogen, phosphorus, and potassium contents in the leaves of different plants using laboratory-based visible and near-infrared reflectance spectroscopy: comparison of partial least-square regression and support vector machine regression methods. *International journal of remote sensing* **34**, 2502–2518 (2013).
15. Ariana, D. P., Lu, R. & Guyer, D. E. Near-infrared hyperspectral reflectance imaging for detection of bruises on pickling cucumbers. *Computers and electronics in agriculture* **53**, 60–70 (2006).
16. Gualtieri, J. A., Chettri, S. R., Cromp, R. F., & Johnson, L. F. Support vector machine classifiers as applied to AVIRIS data. In Proc. Eighth JPL Airborne Geoscience Workshop (1999, February).
17. Baumgardner, M. F., Biehl, L. L., & Landgrebe, D. A. 220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3. Purdue University Research Repository. https://doi.org/10.4231/R7RX991C (2015).
18. Steddom, K., Bredehoeft, M., Khan, M. & Rush, C. Comparison of visual and multispectral radiometric disease evaluations of Cercospora leaf spot of sugar beet. *Plant Disease* **89**, 153–158 (2005).
19. Ahmad, M., Khan, A. M., Hussain, R., Protasov, S., Chow, F., & Khattak, A. M. Unsupervised geometrical feature learning from hyperspectral data. In Computational Intelligence (SSCI), 2016 IEEE Symposium Series on (pp. 1–6). IEEE (2016, December).
20. Zhu, H. *et al.* Hyperspectral imaging for presymptomatic detection of tobacco disease with successive projections algorithm and machine-learning classifiers. *Scientific Reports* **7**, 4125 (2017).
21. Ahmad, M., Khan, A. M. & Hussain, R. Graph-based spatial–spectral feature learning for hyperspectral image classification. *IET Image Processing* **11**, 1310–1316 (2017).
22. Mou, L., Ghamisi, P. & Zhu, X. X. Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **56**, 391–406 (2018).
23. Ahmad, M. *et al.* Fuzziness-based active learning framework to enhance hyperspectral image classification performance for discriminative and generative classifiers. *PloS one* **13**, e0188996 (2018).
24. Yang, S. *et al.* Semisupervised dual-geometric subspace projection for dimensionality reduction of hyperspectral image data. *IEEE Transactions on Geoscience and Remote Sensing* **52**, 3587–3593 (2014).
25. Cariou, C., Chehdi, K. & Le Moan, S. BandClust: An unsupervised band reduction method for hyperspectral remote sensing. *IEEE Geoscience and Remote Sensing Letters* **8**, 565–569 (2011).
26. Ahmad, M., Haq, D. I. U., Mushtaq, Q. & Sohaib, M. A new statistical approach for band clustering and band selection using K-means clustering. *International Journal of Engineering and Technology* **3**, 606–614 (2011).
27. Dy, J. G. & Brodley, C. E. Feature selection for unsupervised learning. *Journal of machine learning research* **5**, 845–889 (2004).
28. Ahmad, M. *et al.* Segmented and non-segmented stacked denoising autoencoder for hyperspectral band reduction. *Optik* **180**, 370–378 (2019).
29. Liu, Z.-Y., Shi, J.-J., Zhang, L.-W. & Huang, J.-F. Discrimination of rice panicles by hyperspectral reflectance data based on principal component analysis and support vector classification. *Journal of Zhejiang University SCIENCE B* **11**, 71–78 (2010).
30. Schor, N. *et al.* Development of a robotic detection system for greenhouse pepper plant diseases. *Precision agriculture* **18**, 394–409 (2017).
31. Cox, T. F. & Cox, M. A. *Multidimensional scaling*. (Chapman and hall/CRC, 2000).
32. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science* **290**, 2319–2323 (2000).
33. Moghadam, P., Ward, D., Goan, E., Jayawardena, S., Sikka, P., & Hernandez, E. Plant Disease Detection Using Hyperspectral Imaging. In Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on (pp. 1–8). IEEE (2017, November).
34. Xie, F., Li, F., Lei, C., Yang, J. & Zhang, Y. Unsupervised band selection based on artificial bee colony algorithm for hyperspectral image classification. *Applied Soft Computing* (2018).
35. Du, Q. & Yang, H. Similarity-Based Unsupervised Band Selection for Hyperspectral Image Analysis. *IEEE Geosci. Remote Sensing Lett.* **5**, 564–568 (2008).
36. Jia, S., Tang, G., Zhu, J. & Li, Q. A novel ranking-based clustering approach for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing* **54**, 88–102 (2016).
37. Chang, C.-I., Du, Q., Sun, T.-L. & Althouse, M. L. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE transactions on geoscience and remote sensing* **37**, 2631–2641 (1999).
38. Mladenić, D. In *Subspace, latent structure and feature selection* 84–102 (Springer, 2006).
39. Serpico, S. B. & Bruzzone, L. *A new search algorithm for feature selection in hyperspectral remote sensing images*. (University of Trento, 2001).
40. Goodfellow, I. *et al.* Generative adversarial nets. In Advances in neural information processing systems (pp. 2672–2680) (2014).
41. Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097–1105) (2012).
42. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436 (2015).
43. Arjovsky, M., Chintala, S., & Bottou, L. Wasserstein generative adversarial networks. In International Conference on Machine Learning (pp. 214–223) (2017, July).
44. Odena, A., Olah, C. & Shlens, J. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585* (2016).
45. Zhu, L., Chen, Y., Ghamisi, P. & Benediktsson, J. A. Generative Adversarial Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* (2018).
46. Acquarelli, J. *et al.* Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica chimica acta* **954**, 22–31 (2017).
47. Malek, S., Melgani, F. & Bazi, Y. One-dimensional convolutional neural networks for spectroscopic signal regression. *Journal of Chemometrics* (2017).
48. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern classification*. (John Wiley & Sons, 2012).
49. Nigsch, F. *et al.* Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *Journal of chemical information and modeling* **46**, 2412–2422 (2006).
50. Zhu, B. *et al.* Walnut shell and meat differentiation using fluorescence hyperspectral imagery with ICA-kNN optimal wavelength selection. *Sensing and Instrumentation for Food Quality and Safety* **1**, 123–131 (2007).
51. Del Fiore, A. *et al.* Early detection of toxigenic fungi on maize by hyperspectral imaging analysis. *International journal of food microbiology* **144**, 64–71 (2010).

52. Zhang, X., Liu, F., He, Y. & Gong, X. Detecting macronutrients content and distribution in oilseed rape leaves based on hyperspectral imaging. *Biosystems engineering* **115**, 56–65 (2013).
53. Weber, V. *et al.* Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. *Field Crops Research* **128**, 82–90 (2012).
54. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
55. Chollet, F. Keras: The python deep learning library. *Astrophysics Source Code Library* (2018).
56. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).

Acknowledgements

Author Contributions

Dongyi proposed and implemented the whole analytical model, wrote the main manuscripts and prepared figures. Robert participated into the algorithm design. Maxwell and Gary contributed to data preparation and revised the manuscript. Avital and Shimon designed and implemented the hyperspectral imaging station and the image dataset. Yang designed the whole experiment and gave suggestions for this work. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.