CellPress

Feature Review

Machine Learning for High-Throughput Stress Phenotyping in Plants

Arti Singh,[1,*] Baskar Ganapathysubramanian,[2] Asheesh Kumar Singh,[1] and Soumik Sarkar[2]

Advances in automated and high-throughput imaging technologies have resulted in a deluge of high-resolution images and sensor data of plants. However, extracting patterns and features from this large corpus of data requires the use of machine learning (ML) tools to enable data assimilation and feature identification for stress phenotyping. Four stages of the decision cycle in plant stress phenotyping and plant breeding activities where different ML approaches can be deployed are (i) identification, (ii) classification, (iii) quantification, and (iv) prediction (ICQP). We provide here a comprehensive overview and user-friendly taxonomy of ML tools to enable the plant community to correctly and easily apply the appropriate ML tools and best-practice guidelines for various biotic and abiotic stress traits.

Plant Stress Phenotyping in Agriculture

To meet the future demand of food, feed, fiber, and fuel, crop production must be doubled by 2050[i] Crop yields are limited inherently by plant stresses (biotic and abiotic), and plant breeders have protected yield from plant stress losses by incorporating resistance genes and developing more climatically-resilient cultivars. Plant breeders and researchers rely on plant phenotyping for accurate and precise trait collection and use of genetic tools to achieve their research goals. Plant phenotyping is defined as the application of methodologies and protocols to measure a specific trait, ranging from the cellular level to the whole plant or canopy level, related to plant structure and function [1]. Agriculture research programs phenotype large populations for several traits throughout the crop growth cycle. This challenge to phenotype multiple traits and large populations is exacerbated by the necessity of sampling multiple environments and growing replicated trials. Until recently, traditional methods of phenotyping have not kept pace with the available high-throughput genotyping tools. The bottleneck in phenotyping has driven intense efforts by the scientific community of agriculture researchers and engineers to adapt newer technologies in field phenotyping. A classic example is high-throughput phenotyping (HTP), which has unlocked new prospects for non-destructive field-based phenotyping in plants for a large number of traits including physiological, biotic (includes living factors such as fungi, bacteria, virus, insects, parasites, and weeds, etc.) and abiotic (includes non-living factors such as drought, flood, nutrient deficiency, and other environmental factors) stress traits [2,3]. Both ground and aerial HTP platforms, equipped with multiple sensors are being used in agriculture to measure multiple plant traits at varying growth stages rapidly, precisely, and accurately (Figure 1A, Key Figure). Examples of these HTP platforms include deployment in cotton (*Gossypium hirsutum* L.) [4], triticale (× *Triticosecale* Wittmack L.) [5], and maize (*Zea mays* L.) [6]. Recent advances in sensors for imaging plants [7,8], ranging from remote sensing including spectroradiometry [9], Light Detection and Ranging (LIDAR) [10], visible to far-infrared

Trends

High-throughput phenotyping (HTP) has unlocked new prospects for non-destructive field-based phenotyping. Autonomous, semi-autonomous, or manual platforms equipped with single or multiple sensors collect spatial and temporal data, resulting in massive amounts of data for analysis and storage.

The enormous volume, variety, and velocity of HTP data generated by such platforms make it a 'big data' problem. Big data generated by these near real-time platforms must be efficiently archived and retrieved for analysis. The analysis and interpretation of these large datasets is quite challenging.

Sophisticated data collection, storage, and processing are becoming ubiquitous, and newer areas of application are emerging constantly. One such relatively new domain is plant stress analytics.

ML algorithms are a very promising approach for faster, more efficient, and better data analytics. ML being inherently multidisciplinary draws inspiration and utilizes concepts from probability theory, statistics, decision theory, optimization, and visualization.

Most current applications of ML tools in plant sciences have focused on using a limited set of ML tools (SVM, ANN). A good understanding of which, when, and how various ML tools can be applied will be very informative to the plant community to leverage these ML tools.

[1]Department of Agronomy, Iowa State University, Ames, IA, USA

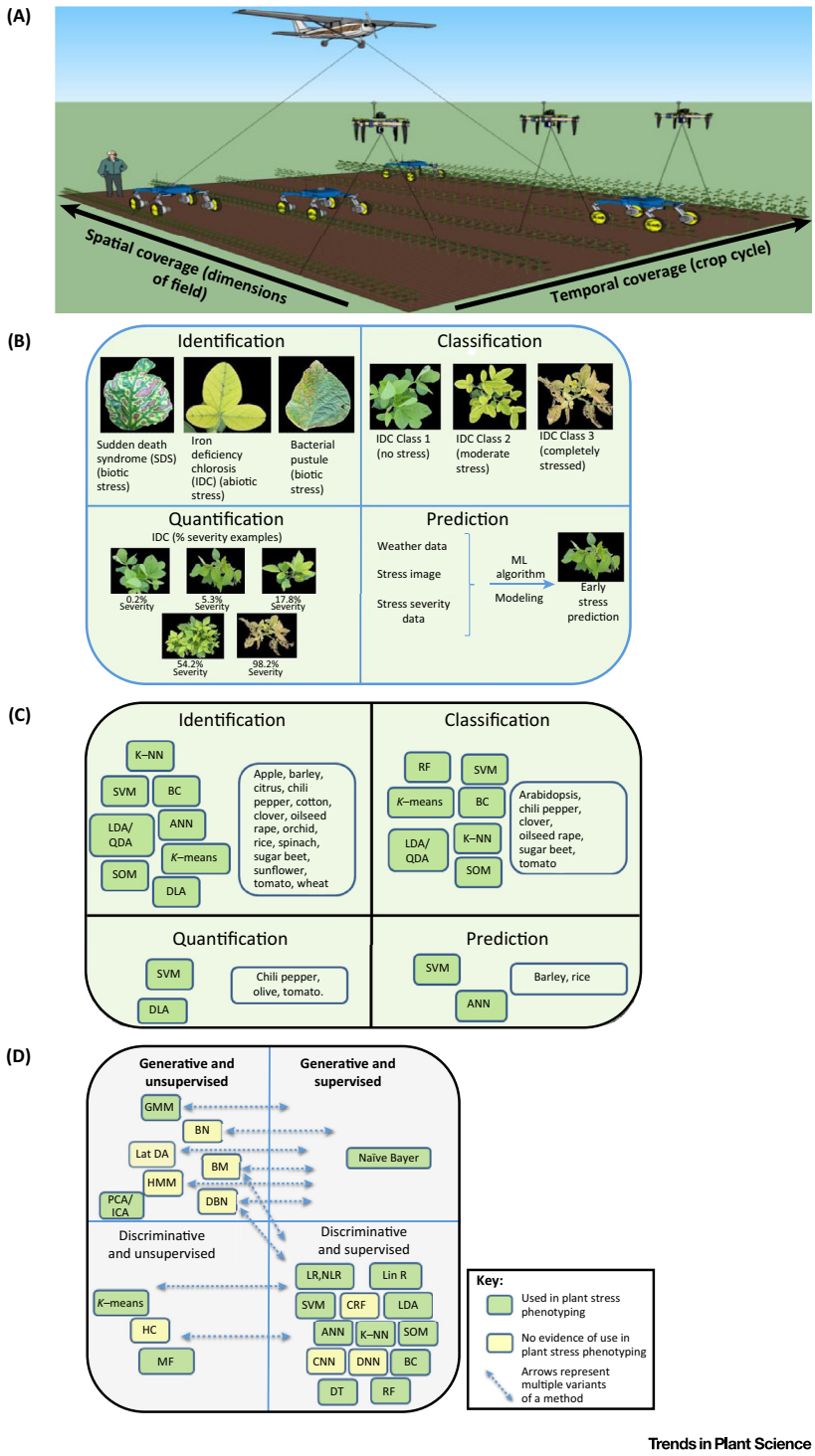
CellPress

Key Figure

Machine Learning (ML) Tools for High-Throughput Stress Phenotyping

[2]Department of Mechanical Engineering, Iowa State University, Ames, IA, USA

*Correspondence: arti@iastate.edu (A. Singh).



Trends in Plant Science

[11], hyperspectral [12–15], thermal [16,17], fluorescence [16,18], and 3D laser scanning [19] to trichromatic (RGB) [20] imaging in conjunction with advanced autonomous vehicles, have truly opened up the possibility of high-throughput stress phenotyping (HTSP). Autonomous platforms such as unmanned aerial vehicles (UAVs) [21] and ground robots [22] equipped with multiple sensors can take pictures in near real-time of the entire experimental plot several times per day, or over the entire season from germination to maturity, resulting in massive amounts of data for analysis and storage. Making sense of all these collected data can be done effectively using ML tools (see the supplemental information online). It is important to emphasize that for ML tools data can be collected from complex, integrated imaging platforms or from simple(r) methods such as crowd-sourced cell phone images.

The objectives of this review are (i) to give an overview of work done in the field of plant stress phenotyping using ML in identification, classification, quantification, and prediction (ICQP); (ii) to give an overview of general issues in ML to develop a strategy for wider application and adaptability in agriculture; (iii) to contribute to a technical framework for the application of ML in plant breeding to solve practical problems, especially plant stress phenotyping using digital imaging; and (iv) to discuss the advantages and disadvantages of various ML algorithms in stress phenotyping with the aim of enabling further exploration of these tools for facilitating practical use in plant breeding.

Phenotyping Data and ML

The enormous volume, variety, velocity, and veracity of imaging and remote-sensing data generated by such real-time platforms represent a 'big data' problem. The data generated by these near real-time platforms must be efficiently archived and retrieved for analysis. Although the analysis and interpretation of such (image-based) big data are challenging, the ensuing possibilities that can impact on agricultural production make it a promising approach for HTP and HTSP. ML approaches present a scalable, modular strategy for data analysis, especially for the new application domain of 'plant stress analytics'. Recent studies on HTSP using images obtained from UAV-based platforms to detect weeds in wheat (*Triticum aestivum* L.) [23], maize [21], and sunflower (*Helianthus annuus* L.) [24] using ML algorithms have paved a new path for better stress management practices on spatial and temporal basis.

ML is an inherently multidisciplinary approach to data analysis that draws inspiration, and borrows heavily, from probability theory, statistics, decision theory, visualization, and optimization. ML approaches are typically useful in situations where large amounts of data are available, relating inputs (e.g., image data) to output quantities of interest (e.g., stress phenotypes). One of the major advantages of using ML approaches for plant breeders, pathologists, physiologists, and biologists is the opportunity to search large datasets to discover patterns and govern discovery by simultaneously looking at a combination of factors instead of analyzing each feature (trait) individually. This was previously a major bottleneck because the high dimensionality of individual images (coupled with the huge number of such images) makes them extremely difficult to analyze through classical techniques. Another key challenge is that the underlying

Figure 1. (A) High-throughput stress phenotyping in soybean field at various growth stages and at different heights using aircraft, UAV, and UGV. (B) Identification, classification, quantification, and prediction (ICQP) of plant diseases in soybean. (C) ML algorithms used in ICQP of plant stresses. (D) Classification of ML algorithms into generative and discriminative. Abbreviations: ANN, artificial neural network; BC, Bayes classifier; BN, Bayesian network; BM, Boltzmann machine; CRF, conditional random field; CNN, convolutional neural network; DT, decision tree; DNN, deep neural network; GMM, Gaussian mixture models; GP, Gaussian process; HMM, hidden Markov model; HC, hierarchical clustering, ICA, independent component analysis; K-MC, *K*-means clustering; K-NN, *k*-nearest neighbor classifier; Lat DA, latent Dirichlet allocation, LDA, linear discriminant analysis; Lin R, linear regression; LR, logistic regression; MF, matrix factorization; NB, naïve Bayes; NLR, nonlinear regression; PCA, principal component analysis; RF, random forests; SOM, self-organizing map; SVM, support vector machine; UAV, unmanned aerial vehicle; UGV, unmanned ground vehicle.

processes for linking the inputs to the outputs are too complex to model mathematically. This is particularly the case for plant stress phenotyping, where it is challenging to efficiently model the holistic effect of genetic, agronomic, economic, meteorological, and human factor inputs on stress and, ultimately, yield.

ML methods have been applied with spectacular results to similar problems [25] previously presumed to be impossible to model. Examples include numerous success stories in various domains ranging from computer vision (e.g., face recognition), speech processing (e.g., Google voice, Apple's Siri) and natural language processing (e.g., IBM Watson), consumer predictive analytics (e.g., Netflix movie recommender system) to bioinformatics (e.g., personalized genomics, drug design, and genome annotation) [26], cell biology [27], and disease tissue classification in medicine [28,29]. The success of ML tools is attributed to their ability to identify a hierarchy of features and generalized trends from available data. These tools have also proved particularly adept at integrating disparate and often redundant data to draw coherent (and often non-intuitive) patterns for identification and quantification. Finally, current progress in ML has resulted in scalable, robust, and flexible software tools (R packages, Matlab toolboxes, and software packages such as Theano, Caffe for archetypal ML algorithms) [30–33] that make the application of ML to disparate disciplines straightforward. The advances have percolated into agricultural research, where there has been an increasing research effort to apply ML approaches in diverse species, such as horticultural crops and forest tree species. This has been driven in part by the increased investment by commercial companies and the decreasing cost of imaging/sensor platforms.

ML-Enabled HTSP

In light of these developments, it is clear that ML-enabled HTSP will benefit plant breeders, physiologists, entomologists, pathologists, extension workers, and farmers by allowing screening of different stresses in an accurate, precise, and speedy manner (Table 1). This will also directly enable the acceleration of the gene discovery process as well as the introduction of novel selection protocols for complex quantitative traits such as biotic and abiotic stresses and yield. ML-enabled HTSP will also enhance our understanding of pathogen–plant interactions [34] as well as the interaction of plants with other stresses.

What is ML?

ML refers to a group of computerized modeling approaches that can learn patterns from the data so as to make automatic decisions without programming explicit rules. The main idea of ML is to effectively utilize experiences or example scenarios to discover underlying structures, similarities, or dissimilarities present in data to explain or classify a new experience or an example scenario properly. A key ability of ML tools is their ability to generalize trends and/or patterns from available data.

There are a large number of choices of ML tools. It is important for an application expert to make a judicious choice on a specific ML method to deploy for his/her specific problem. We advocate that the practitioner should (carefully) down-select from the plethora of ML choices based on the type and amount of available data and problem formulation. In the context of plant stress phenotyping, we identify four distinct classes of problem formulation: (i) identification/detection, (ii) classification, (iii) quantification/estimation, and (v) prediction (Figure 1B,C). Furthermore, preprocessing steps such as dimension reduction, clustering, and segmentation can also be crucial for successful decision-making.

Structure of a ML Process

Typically, a large fraction of the dataset, known as the ‘training dataset’, that represents the entire population is used for calibrating the model. The remaining dataset is used to test the

Table 1. Examples of ML Approaches in Plant Species for Stress Phenotyping

ML Algorithm Application in HTSP	ML Algorithm Type	Sensor	Plant Species	Trait(s) Phenotyped	Stress Type	Refs
Identification	SVM with a linear kernel	Thermal and stereo visible light	Tomato (*Solanum lycopersicum* L.)	Powdery mildew	Disease	[31]
Identification	SAM	Remote sensing	Sugar beet (*Beta vulgaris* L.)	*Heterodera schachtii* and *Rhizoctonia solani*	Pest and disease	[44]
Identification	None Preprocessing via segmentation	Kinect RGB depth images	Apple (*Malus domestica* Borkh.)	Apple scab	Disease	[70]
Identification	SVM and Gaussian processes classifier (GPC)	Visible and thermal images	Spinach (*Spinacia oleracea* L.)	Drought/water stress	Abiotic stress	[71]
Identification	Bayes factor and DAR	Hyperspectral images	Barley (*Hordeum vulgare* L.)	Rust, net blotch, and powdery mildew	Disease	[11]
Identification	SVM	Fluorescence imaging spectroscopy	Citrus [*Citrus sinensis* (L.) Osbeck]	Huanglongbing (HLB)	Disease	[36]
Identification	OBIA-based classification	UAV-based RGB images and multispectral image	Sunflower (*Helianthus annuus* L.)	Weed	Biotic stress	[21]
Identification	None Preprocessing via segmentation	RGB images	Cotton (*Gossypium hirsutum* L.)	Southern green stink bug, bacterial angular and *Ascohyta* blight	Disease and insect	[39]
Identification	SVM, linear kernel, quadratic kernel (QP), radial basis function (RBF), multilayer perceptron (MLP), and polynomial kernel	RGB images	Tomato	Tomato yellow leaf curl virus and tomato yellow leaf curl disease	Disease	[40]
Identification	ANN variant	RGB images	Orchid (*Phalaenopsis*)	Bacterial soft rot, *Phythopthora* black rot, bacterial brown spot	Disease	[42]
Identification	SVM	UAV- and aircraft-based sensors	Citrus	Huanglongbing (HLB)	Disease	[37]
Identification	KNN, quadratic discriminant analysis (QDA), and linear discriminant analysis (LDA)	Spectroradiometer	Citrus	Huanglongbing (HLB)	Disease	[43]
Identification	SViM	Hyperspectral	Tomato	Water stress	Drought	[72]
Identification	Gaussian mixture model	RGB images	Wheat (*Triticum aestivum* L.)	Wheat streak mosaic virus	Disease	[45]
Identification	SVM variant	Scanned images	Rice (*Oryza sativa* L.)	Nitrogen, phosphorus, and potassium (NPK) stress	Nutrient deficiency	[73]
Identification and classification	HBBE, MLPNNs, LDA	CCD images	Chili pepper (*Capsicum annuum* L.)	Aflatoxins	Toxic metabolites	[52]
Identification and classification	SVM	Hyperspectral reflectance	Sugar beet	*Cercospora* leaf spot, sugar beet rust, and powdery mildew	Disease	[38]
Identification and classification	Linear discriminant analysis (LDA) and *K*-means	RGB images	Clover (*Trifolium subterraneum* L.)	Ozone	Pollution	[53]
		RGB images	Tomato	Powdery mildew	Disease	[20]

Table 1. (continued)

ML Algorithm Application in HTSP	ML Algorithm Type	Sensor	Plant Species	Trait(s) Phenotyped	Stress Type	Refs
Identification and classification	Self-organizing map (SOM)					
Identification and classification	Bayesian classifier	RGB images	Tomato	Powdery mildew	Disease	[20]
Identification and classification	Naïve Bayes (NB), simple logistic (SL), LibSVM (SVM), LibLINEAR (LINE), MLP (BNN), functional trees (FT), random forests (RF), classifier for generating a grafted C4.5 decision tree (J48)	Hyperspectral images	Oilseed rape (*Brassica napus* L.)	*Alternaria alternata, Alternaria brassicae, Alternaria brassicicola, and Alternaria dauci*	Disease	[48]
Classification	SVM	RGB images	Arabidopsis (*Arabidopsis thaliana* L.)	*Salmonella* bacteria	Disease	[49]
Classification	Random forest and SVM spatial matching kernel	–	–	Stonefly larvae	Insect	[51]
Classification	*k*-NN and Bayesian classifier	Fusion of RGB and multispectral image	Sugar beet	*Uromyces betae, Cercospora beticola*	Disease	[47]
Classification	Bayesian classifier	RGB images	Arabidopsis	*Salmonella* bacteria	Disease	[50]
Quantification	SVM	RGB images and spectral reflectance	Tomato	Leaf miner	Insect	[57]
Quantification	None Preprocessing via segmentation	RGB images	Chili pepper	Yellow vein virus	Disease	[56]
Quantification	SVM and LDA	Hyperspectral sensor and thermal images	Olive	*Verticillium dahliae*	Disease	[58]
Prediction	Dirichlet aggregation regression (DAR)	Hyperspectral images	Barley	Drought	Abiotic stress	[46]
Prediction	SVM, generalized regression neural network (GRNN)	Manual severity rating	Rice	Rice blast	Disease	[59]
Prediction	SVM	Hyperspectral images	Barley	Drought	Abiotic stress	[60]

calibrated model, and is termed the 'testing dataset'. The next step after training is to validate the learnt model on a new set of data (from different or same population). Once the accuracy and precision of the model is high enough, it can be used on a routine basis to identify, classify, quantify and predict particular stress features.

We discuss this process in greater detail next, particularly focusing on clarifying the choice of ML tools to use based on two different points of view, namely the learning process (how are features learnt) and the modeling objective (what is being learnt).

The Learning Process – Supervised versus Unsupervised

The first point of view concerns whether or not the ML model is provided with the labels to the data it uses for model training. Here, a label is a trait (such as diseased plant or type of crop plant) that is associated with an image. For example, suppose the objective of a ML model is to distinguish between maize, soybean [*Glycine max* (L.) Merr.], and sunflower plants after

collecting a large number of images. If the model is trained with a set of such images where each is labeled as soybean, maize or sunflower, the learning process is termed 'supervised'. Alternatively, if the training images are provided without any label, the learning process is 'unsupervised'. Note that although some ML tools can be trained either in a supervised or an unsupervised manner, some can be trained in both ways. With the supervised learning process, a model essentially tries to learn a map between the input dataset and the corresponding output labels. For the given example, it means that a model trained in a supervised manner learns how to map all the soybean examples to the soybean label, and so on.

Support vector machines (SVM) and regular artificial neural networks (ANN) are among the prominent examples of supervised schemes. By contrast, an unsupervised method does not have specific output labels associated with input images. It identifies structures or features present in the images, such as the presence of tassels in maize, or the presence of trifoliate leaves in soybean. In many cases, the features identified in the unsupervised process may not be meaningful to a human user. Various clustering, mixture models, and dimension-reduction techniques fall under this category. Often ML models can be developed by using partly labeled data. This training scheme is known as 'semi-supervised' learning.

The Modeling Objective – Generative versus Discriminative

A second point of view for ML method categorization is the modeling objective: whether the model being trained is to distinguish between two different data patterns or to be able to learn and generate similar patterns synthetically. In the context of the previous example, a maize and soybean plant can be distinguished based on the presence or absence of tassels or pods. A model developed with only this difference would be termed a 'discriminative model'. Naturally, many supervised methods such as SVM and ANN fall under this category. However, such a model does not learn other features of the objects. Thus, a discriminative model that distinguishes a soybean plant from a maize plant based on the presence of tassels fails to differentiate between an image of a maize plant and that of sorghum [*Sorghum bicolor* (L.) Moench] plant. Therefore, a discriminative model is typically built for a predetermined specific task. On the other hand, a model that captures the overall data pattern such that it is able to generate synthetic images is known as a 'generative model'. It follows that a single generative model can be useful for many decision tasks at the same time. Mixture models, hidden Markov models, and Boltzmann machines are prominent examples of generative models.

Given a large volume of data, discriminative models usually perform better than generative models, especially for classification tasks, such as distinguishing images of soybean plants from a maize plant. However, generative models can achieve slightly better performance compared to discriminative ones with low training data volume. In addition, generative models tend to be more robust to overfitting issues (i.e., a model learns training data too well and performs very poorly for unseen test data).

Figure 1D schematically illustrates the categorization of several ML methods into classes of generative versus discriminative, and supervised versus unsupervised. Although it seems that supervised models are more likely to be discriminative in nature, and generative models should not need output labels for training, there are examples of unsupervised discriminative models and supervised generative models. For example, the widely popular *K*-means clustering technique usually follows an unsupervised training scheme. Even so, it is not possible to generate new data examples reliably using only the cluster centers and groupings. Most of the unsupervised generative techniques can be learned in a supervised manner when the target class information is incorporated as one of the data features. Apart from such variations (as shown in Figure 1D), many of the ML tools have minor variations based on particular aspects such as underlying model structure (e.g., linear vs nonlinear) and training algorithm.

The Role of Preprocessing in ML-HTSP

The crucial step for the successful deployment of ML methods is careful preprocessing of the image data. There are multiple examples where a careful choice of preprocessing of the collected image datasets has resulted in substantial improvements in ML performance. Preprocessing can vary from very simple operations including image cropping, contrast enhancement, and removal of background to significantly more complex operations such as clustering and dimensionality reduction using principal component analysis (PCA). However, the overarching principle of preprocessing the data is 'concentration of information'. That is, the original datasets may contain a large quantity of unnecessary or conflicting information that can result in poor ML performance. By preprocessing, the signal-to-noise ratio (ratio of useful to useless information) is improved. This directly enhances the ability of the ML model to easily recognize useful patterns or trends and to separate the data into appropriate classes.

Preprocessing is the stage where domain knowledge is crucial. The domain expert identifies features in the image that are relevant or important for training the model. For example, the removal of background (soil, dirt, and tags, etc.) from the foreground to identify the plant canopy is generally a crucial step. Following this, a variety of image processing tools can be used to convert these raw datasets into a more relevant dataset that contains the extracted features. Examples of preprocessing operations include: (i) segmentation of images; (ii) contrast enhancement to detect edges; (iii) thresholding images into binary data; (iv) converting one image format into other [RGB to greyscale; RGB to hue saturation value (HSV)]; (v) de-noising images using filters [band-pass, low-pass, fast Fourier transform (FFT)]; (vi) extracting features at different scales using image transforms (FFT, wavelet transforms, Haar transforms, Hough transforms, Radon transforms); (vii) pixel-based classification; (viii) clustering of images into classes; and (ix) dimensionality reduction of images. Several of the ML tools discussed here can in fact be applied to these preprocessing stages.

There are several examples of the utility of such preprocessing steps. For early site-specific weed management (ESSWM) in wheat, a UAV-based platform was equipped with a visible-range camera to capture ultra-high resolution images. Otsu's thresholding method was then used to differentiate wheat crop plants from weeds [23]. In sunflower, a quadrocopter equipped with RGB and multispectral sensors was used to detect weeds to optimize herbicide application. A pixel-based classification approach was used for weed identification. Similarly, image segmentation [21,24,35] followed by automatic object-based image analysis (OBIA) was used in sunflower and maize to identify spatial and spectrally consistent objects.

ML Approaches to ICQP

Identification of Stress

Identification methods involve detection of a specific stress amid other potential stresses in the field. Here, preprocessing of image data is crucial. This is especially important for applications involving high-throughput imaging, where plant stressors (such as weeds, nutrient, disease, and insects) must be automatically identified. In the past, ML methods such as SVM, neural networks (NNs), kernel methods, and instance-based approaches have been used to detect various stresses. The SVM method has been used successfully in a variety of scenarios for stress identification in plants.

SVM Methods

SVM methods have been applied to a variety of plants for disease and stress identification. They are deployed in the citrus industry to contain citrus greening (caused by phloem-limiting bacteria), also known as Huanglongbing (HLB). Early identification is crucial for effectively controlling the spread of HLB. ML was used for the identification and estimation of HLB based on fluorescence imaging spectroscopy data of leaf samples [36]. The images were

preprocessed (segmented) and the extracted features were used as an input to SVM. In another example, two aerial imaging platforms were compared for the identification of HLB diseases in an infected citrus orchard. An aircraft sensing platform equipped to capture hyperspectral images and a UAV platform equipped with a multiband imaging camera were used, and images obtained from both platforms were segmented and piped to a SVM classifier. Results showed that a (non-linear) SVM with kernel worked better than (linear) SVM, LDA, and QDA [37]. Similarly, in sugar beet (*Beta vulgaris* L.), early identification of three diseases, *Cercospora* leaf spot, leaf rust, and powdery mildew, was performed using SVM with a radial basis function as kernel [38]. The same idea was deployed in cotton to identify damage by green stink bug, bacterial angular blight, and *Ascochyta* blight using a SVM classifier [39], as well as in tomato (*Solanum lycopersicum* L.) to identify viruses: tomato yellow leaf curl virus and tomato yellow leaf curl disease [40].

SVM (combined with a Gaussian process classifier) was used for the automatic identification of soil moisture stress using remote visible and thermal images in spinach canopies, where the efficacy of using a combination of methods was explored. In another study, visible and thermal imaging were combined with depth information to automatically identify powdery mildew of tomato plants at an early growth stage using a SVM classifier kernel [31]. Variants of the SVM method have also proven to be useful for identification. In rice (*Oryza sativa* L.], the hierarchical identification of nutrient deficiency symptoms using scanned images was carried out using support vector feature selection (SVFS). Hierarchical identification was able to detect the NPK (nitrogen, phosphorus, and potassium) stress effectively with enhanced identification accuracy [41].

Artificial Neural Nets (ANN) and Variants

Simple image processing, followed by ANN, was used for detecting and identifying three orchid seedling diseases: bacterial soft rot, bacterial brown spot, and *Phytopthora* black rot [42].

Discriminant Analysis

Spectral reflectance data were recorded from healthy and HLB-infected citrus leaves, and various classifier methods (discriminative and supervised methods) were used to identify infected leaves using quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), and *k*-nearest neighbor (*k*-NN) and soft independent modeling of class analogy (SIMCA) [43]. In sugar beets, hyperspectral canopy reflectance was used to identify stress symptoms instigated by beet cyst nematode and *Rhizoctonia* crown and root rot [44].

Gaussian Mixture Models

Preprocessing followed by the application of Gaussian mixture models has been used to identify both biotic and abiotic stresses [45]. Canopy images were taken in a stressed wheat field, at differential levels of irrigation (abiotic stress) and inoculated with wheat streak mosaic virus (biotic stress).

K-means Clustering and Variants

Water and nutrient stress at the crop canopy level have been studied using hyperspectral images that were analyzed using simplex volume maximization (SiVM), an unsupervised clustering approach. The SiVM algorithm was found to be adept at detecting drought stress using hyperspectral images four days before the appearance of symptoms visible to the human eye. The unsupervised nature of SiVM also provided useful new insights into the data [11,25].

Dimensionality Reduction and Clustering

The automatic extraction of hyperspectral signatures using the Bayes factor algorithm with Dirichlet-aggregation regression (DAR) resulted in rapid, reliable, and data-driven approach for

phenotyping plants for both biotic and abiotic stresses [11]. In barley (*Hordeum vulgare* L.), the DAR algorithm was used for the presymptomatic prediction of drought stress at an early stage using hyperspectral images [46].

Self-Organizing Map (SOM) and Bayes Classifier

Both supervised (Bayesian classifier) and unsupervised (SOM) ML approaches were used for segmenting diseased tomato plants. Diseased regions on the tomato canopy were identified using the preprocessed tomato images by a SOM model [20].

Classification of Stress

Classification is an extension of identification; however, instead of identifying a particular stress amid different stresses, a classifier is used to classify stress into labeled classes on the basis of stress symptoms and signatures. Classification methods often include a preprocessing step, usually a segmentation step followed by the extraction of features that feed into some type of classifier. For instance, drought stress can be classified into: no stress, moderate stress, and heavy stress. Examples from various crops are presented below illustrating the use of classification algorithms in various stresses.

ANN

A combination of *K*-means clustering followed by ANN, using fused information from RGB and multispectral camera imaging, was used to develop a 3D model of leaves to differentiate the two sugar beet diseases, *Cercospora beticola* and *Uromyces betae* [47]. In another example, thermal and hyperspectral imaging were analyzed to identify and classify symptoms caused by three *Alternaria* species in the oilseed brassica. The majority of the available classifiers present in Weka (useful data-mining software in java) were tested, and eight classifiers were used to make comparisons. The back-propagation neural networks (BNNs) classification model displayed the highest prediction accuracy for classifying *Alternaria* species [48].

SVM

A SVM-based method was used to classify healthy and unhealthy *Arabidopsis* plants on the basis of symptoms instigated by colonization of the human pathogen *Salmonella* Typhimurium [49]. The SVM model was comprehensively tested on 1200 individual *Arabidopsis* plants with positive results [50]. In further work, classification error was reduced using SVMs compared to other ML methods such as decision trees and ANNs. The classification accuracy of three sugar beet diseases increased with increasing disease severity. This is especially true when hyperspectral reflectance-based vegetation indices were used with SVM for the automatic classification of disease severity of three sugar beet diseases (*Cercospora* leaf spot, leaf rust, and powdery mildew). Another example of the usefulness of SVM was the identification and classification of stonefly larvae from images. Haar random forest feature extraction, in combination with SVM classifier, was able to differentiate and classify stonefly larvae from other insect species [51].

LDA, K-Means, and Coupled Methods

ML methods were used to classify the presence of aflatoxins, which are toxic compounds produced by fungus *Aspergillus flavus* and *Aspergillus parasiticus* [52]. Fluorescence (UV illumination) and reflectance (halogen excitations) were used as input data in this case. Preprocessing was performed to extract features using Guyon's SVM-RFE, classical Fisher discriminant power, and PCA. These extracted features were used as inputs to classifiers such as LDA (linear discriminant analysis) and MLP. Another paper classified the RGB images into 'healthy' and 'injured' classes of clover plants that were exposed to varying level of ozone. Various pixel-classifying methods were compared such as LDA, *K*-means clustering, FPM-T2 (fit to a pattern multivariate image analysis combined with T2 statistics), and FPM-RSS (FPM combined with residual sum of squares statistics) [53].

Box 1. Key Take-Away Points for Practitioners

Identification

(i) A large variety of ML methods have been successfully applied to the disease identification problem. This is an area where preprocessing of images will be very useful.

(ii) There are two ways to frame a disease identification problem for ML based on the amount of data available. (a) When statistically significant nominal (healthy) and diseased datasets are available, it is best to use supervised discriminative models that are trained to distinguish between these classes. (b) In the absence of a statistically significant amount of data (or unbalanced data, see next point below), the best approach is to learn the nominal model of the un-diseased plant (which is feasible due to the availability of data of the healthy plant) using unsupervised methods. Then simple outlier detection can be deployed to identify off-nominal (or diseased) cases.

(iii) We strongly encourage practitioners to refrain from training disease identification models using very small datasets. It is especially important to be vigilant against unbalanced data for training where one state (usually, the healthy state) has much more data instances than the other diseased states.

Classification

(i) We encourage the practitioner to ensure the statistical significance of the available dataset.

(ii) Preprocessing of data is crucial, as is using domain knowledge.

(iii) Current trends are to always use supervised methods, which are mostly discriminative.

(iv) Be vigilant against the overfitting problem. A best practice is to always ensure that regularization option is turned on in any ML method.

(v) It is important to note that these methods can only distinguish between known/trained classes. They cannot be deployed to identify unknown/untrained symptoms (so-called extrapolatory mode).

(vi) We encourage practitioners to extract and report detection confidence out of the classifier.

Quantification

(i) There are very few reported applications of ML for stress quantification, and this provides tremendous opportunities for plant scientists and breeders. Most current work formulates the quantification problem as a classification problem with finer resolution.

(ii) Our recommendation is that unsupervised generative models can be very successful for quantification. This can also work with smaller datasets. The basic idea is to learn the nominal (healthy state) model. The severity of the diseased state is quantified as the distance/offset from this nominal state.

(iii) We strongly recommend practitioners to explore the use of Bayesian networks, deep neural nets, and latent Dirichlet allocation for this class of problems.

Prediction

(i) Most prediction applications of ML are limited to early detection of disease onset.

(ii) Current phenotyping approaches can be extended to obtain time-varying traits. Here, more sophisticated models can be used with great utility. Examples include HMM methods, dynamic Bayesian networks, and recurrent neural networks.

Quantification of Stress

There are fewer reports describing the use of ML approaches to quantify stresses, and this provides tremendous opportunities for plant scientists and breeders. Quantification methods are an extension of classification methods where each class is quantified on the basis of stress severity. In case of plant diseases, disease severity [54] can be used to quantify various diseases. For example, rust severity in wheat can be quantified on a scale of 0–100% [55]. Quantification algorithms are usually best preceded by a preprocessing stage to separate foreground from background, edge detection, and contrast enhancement. A good example of the benefit of preprocessing of images for quantification is illustrated in [56] where color mapping (converting from native RGB to other non-native formats such as HSV) followed by segmentation was used to quantify disease severity of PHYVV (pepper huasteco yellow vein virus) on chili pepper (*Capsicum annuum* L.) plants. Examples from various other crops are presented next.

SVM

Leaf miner pest causes major losses in vegetable and ornamental plants. SVM applied to near-infrared spectral reflectance using spectrophotometer and digital leaf images of tomato plants of damaged leaves was used to successfully quantify damage degree (DD) into five levels [57]. Automatic detection method was used to detect and quantify *Verticillium* wilt (VM) on a large scale in olive using hyperspectral sensor and thermal camera mounted on aircraft. The spatial

distribution of VM was assessed at field level and severity classes (0–4 rating scale quantified into 0–100% severity level) predicted by LDA and non-linear SVM methods. Both ML methods showed good results in detecting and quantifying VM severity in olives (*Olea europaea* L.) [58].

Prediction of Stress

The prediction of plant stress at an early stage before it is visible to the human eye has substantial implications for the timely and cost-effective control of stress. There are very few reported activities on using ML for the prediction of stresses, making this the next big frontier for research efforts. This has tremendous implications for prescription farming and precision agriculture.

SVM

In predicting rice blast disease, SVM was used for the development of a weather-based prediction model [59]. The performance of the SVM-based approach was compared to the ANN variants; back-propagation neural network (BPNN) and generalized regression neural network (GRNN), and also to conventional multiple regression. The SVM-based approach outpaced all the three methods in cross-location and cross-year models, indicating their role in early forecasting of plant diseases. Other examples include the prediction of drought-induced stress in barley using hyperspectral images using an SVM variant, ordinal classification approach [60]. In barley for the presymptomatic detection of water stress, DAR algorithm was used on images obtained using hyperspectral camera. The DAR algorithm was efficient in predicting stress before symptoms were visible to human eye [46].

Strategies for the Development of Efficient ML Applications in Plant Breeding

ML methods can play an extensive role in breeding for stress tolerance and for rapid phenotyping. These approaches can be used in decision-making for parent selection to use in

Box 2. Precision Phenotyping Using ML Algorithms in Agriculture: Best Practices

Advances in technology have made HTSP feasible. Appropriate ML tools can be used for all four stages (stress identification, classification, quantification, and prediction). High-throughput image-based phenotyping of plant stress-related traits to complement high-throughput sequencing will assist in finding new genes and quantitative trait loci (QTLs) using linkage mapping and genome-wide association studies (GWAS) together with training genome-wide selection models for various plant stress traits [74]. Specifically, the high-throughput phenotypic information on a particular stress can be used in association with genotypic information by means of QTL [75,76], GWAS [77,78], or expression studies to bridge the genotype–phenotype gap. Relating time-series phenotypic data with time-series gene expression may provide novel insights into cellular mechanisms. Time-series image data [79,80] obtained in experiments using HTP platforms such as UAVs and autonomous rovers (ground robots), integrated into a viable ML pipeline, will allow the study of the time-dependent gene turn-on and turn-off mechanisms and provide insight into the molecular basis of disease resistance. HTP can also be used for phenotyping different fields for the same stress to understand the spatiotemporal difference in the expression of stress. The big advantage of ML algorithms is that plant scientists can use them proficiently in stress identification, classification, quantification, and prediction using tools packaged in the graphical user interface (GUI) without knowing the underlying mathematical and computational complexities. We advocate the following best practices for maximizing the effectiveness of ML tools:

ML for Practitioners

It is useful to identify which type of ML method is best suited by triaging the ML methods based on amount of data available as well as the type of data (labeled vs unlabeled). Figure 1C,D will be beneficial for this purpose.

Discriminative methods work better for labeled and large datasets, while generative methods work better for smaller datasets (both labeled and unlabeled).

It may be worthwhile for practitioners to explore the use of generative unsupervised models as a means to identify latent features from datasets. This is an area that is relatively unexplored, but holds great promise.

Preprocessing is where domain knowledge can be leveraged to improve the signal-to-noise ratio in the data.

For ML methods that are classified as discriminative and supervised, the use of preprocessing before deploying ML tools is essential to substantially improve ML performance.

In cases where it is difficult to translate domain knowledge into feature crafting, unsupervised preprocessing methods should be explored. This may be particularly useful for high-dimensional data such as hyperspectral data.

hybridization schemes, and in generation advancement and selection. An example of an application using ICQP is presented below.

Identification

Use HTP platform [UAV, unmanned ground vehicle (UGV)] for taking images of breeding plots in plant stress nurseries, and yield tests. Perform image processing and apply ML algorithms. ML identifies which stress is present in each breeding plot.

Classification

Use ML approaches to classify stress in each breeding plot. For instance, determine if a genotype is resistant or susceptible to a particular stress. This information can be directly used in a breeding decision scheme to select stress resistance genotypes.

Quantification

Use ML approaches to quantify the stress. For example, quantification may be on a 0–100% scale of expression, such as percent infection severity. Use the information generated by ML approach through quantification to make selections to identify stress-resistant genotypes for further testing or commercialization.

Prediction

Use ML approaches on prior disease and weather data to make early stage prediction on the expression of stresses for breeding and selection decisions.

In addition to plant stresses, advances have been made in analyzing other traits using ML to predict yield [61–63], biomass [64,65], root traits [66,67], and adaptation traits [68].

Concluding Remarks and Outlook

ML tools provide a very powerful framework to assimilate data, and the utility of these tools is especially important considering current progress in HTP approaches that easily generate terabytes of data. Appropriate choice and usage of ML tools is crucial for obtaining the maximum possible benefits of these sophisticated approaches. This review provides a comprehensive overview of ML alongside best practices of using these ML tools to enable stress phenotyping (Boxes 1,2). Using advanced ML tools for plant stress phenotyping is a very new area, with the plant community focusing on a small number of ML methods (such as SVM and ANN). As part of this review, we have identified several future avenues for using ML techniques that show tremendous promise but remain currently unutilized by the phenotyping community (see Outstanding Questions). Furthermore, the concepts discussed here can be applied to data collected across the spectrum of complexity and sophistication (from manually captured camera imaging to automated high-throughput imaging systems), as well as scale (from individual plant to plot to field).

The outlook for ML tools in agriculture is very promising. A key ingredient for successful large-scale application of ML is the seamless integration of data analytics within the data collection and curation pipeline. Such a computational ecosystem (that links data collection, data storage, and curation with ML-based data analytics) will open up tremendous opportunities to accelerate breeding and to solve foundational problems in genomics and predictive phenomics. Promising examples of this vision include the Integrated Analysis Platform (IAP) [69] and the i-Plant nitiative[ii] A crucial catalyst for such advances will be to foster multidisciplinary research teams such that advances in engineering, plant sciences, and informatics can be leveraged in a rational way. This review will enable such teams by providing a common language of communication related to ML tools.

Outstanding Questions

Can causal ML models, deployed on time-series of images, identify visual precursors to enable early detection of stress response?

ML methods lead to data-driven discovery of non-intuitive patterns. How can these hierarchical patterns be visualized and then physiologically interpreted by domain scientists?

Advanced supervised ML methods tend to require labeling a large volume of training data. Can methodologies such as mechanical turking, crowd sourcing, and gamification be used to create a large corpus of community-labeled data?

Future phenotyping efforts will be characterized by heterogeneous (visual/non-visual), multiscale (plant/plot/field), asynchronous data collected by a variety of sensors connected via the internet of things (IOT). How to develop best practices for a robust, scalable, and integrated ML pipeline that enables seamless assimilation?

The ICQP paradigm comprehensively characterizes the spectrum of problems in plant stress phenotyping, thus enabling the design and deployment of appropriate ML tools. How can this paradigm be applied to other problems in plant biology, such as physiogenetic drivers of yield, and soil microbiome–plant root interaction studies?

Acknowledgments

This work was supported by Iowa State University (ISU) and the Iowa Soybean Association. B.G. thanks the ISU Plant Sciences Institute for support. We are grateful to Dr P. Jayashankar and Ms J. Hicks for reviewing this article and Marcus Naik for helping with the development of figures.

Supplemental Information

Supplemental information associated with this article can be found, in the online version, at doi:10.1016/j.tplants.2015.10.015.

Resources

[i] www.globalharvestinitiative.org/index.php/gap-report-gap-index/2013-gap-report/2013-gap-report-digital/

[ii] www.iplantcollaborative.org/

References

1. Ghanem, M.E. *et al.* (2014) Physiological phenotyping of plants for crop improvement. *Trends Plant Sci.* 20, 139–144
2. Deery, D. *et al.* (2014) Proximal remote sensing buggies and potential applications for field-based phenotyping. *Agronomy* 4, 349–379
3. White, J.W. *et al.* (2012) Field-based phenomics for plant genetics research. *Field Crops Res.* 133, 101–112
4. Andrade-Sanchez, P. *et al.* (2013) Development and evaluation of a field-based high-throughput phenotyping platform. *Funct. Plant Biol.* 41, 68–79
5. Busemeyer, L. *et al.* (2013) BreedVision – a multi-sensor platform for non-destructive field-based phenotyping in plant breeding. *Sensors* 13, 2830–2847
6. Liebisch, F. *et al.* (2015) Remote, aerial phenotyping of maize traits with a mobile multi-sensor approach. *Plant Methods* 11, 9
7. Li, L. *et al.* (2014) A review of imaging techniques for plant phenotyping. *Sensors* 14, 20078–20111
8. Mutka, A. and Bart, R. (2015) Image-based phenotyping of plant disease symptoms. *Front. Plant Sci.* 5, 734
9. Mahlein, A. *et al.* (2010) Spectral signatures of sugar beet leaves for the detection and differentiation of diseases. *Precis. Agric.* 11, 413–431
10. Omasa, K. *et al.* (2007) 3D lidar imaging for detecting and understanding plant responses and canopy structure. *J. Exp. Bot.* 58, 881–898
11. Wahabzada, M. *et al.* (2015) Metro maps of plant disease dynamics-automated mining of differences using hyperspectral images. *PLoS ONE* 10, e0116902
12. Bauriegel, E. *et al.* (2011) Early detection of *Fusarium* infection in wheat using hyper-spectral imaging. *Comput. Electron. Agric.* 75, 304–312
13. Mahlein, A-K. *et al.* (2012) Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases. *Plant Methods* 8, 3
14. Berdugo, C.A. *et al.* (2014) Fusion of sensor data for the detection and differentiation of plant diseases in cucumber. *Plant Pathol.* 63, 1344–1356
15. Ashourloo, D. *et al.* (2014) Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements. *Remote Sens.* 6, 5107–5123
16. Calderón, R. *et al.* (2013) High-resolution airborne hyperspectral and thermal imagery for early detection of Verticillium wilt of olive using fluorescence, temperature and narrow-band spectral indices. *Remote Sens. Environ.* 139, 231–245
17. Prashar, A. *et al.* (2013) Infra-red thermography for high throughput field phenotyping in *Solanum tuberosum*. *PLoS ONE* 8, e65816
18. Rousseau, C. *et al.* (2013) High throughput quantitative phenotyping of plant resistance using chlorophyll fluorescence image analysis. *Plant Methods* 9, 17
19. Paulus, S. *et al.* (2014) Automated analysis of barley organs using 3D laser scanning: an approach for high throughput phenotyping. *Sensors* 14, 12670–12686
20. Hernandez-Rabadan, D.L. *et al.* (2014) Integrating SOMs and a Bayesian classifier for segmenting diseased plants in uncontrolled environments. *Sci. World J.* 2014, 214674
21. Pena, J.M. *et al.* (2015) Quantifying efficacy and limits of unmanned aerial vehicle (UAV) technology for weed seedling detection as affected by sensor resolution. *Sensors* 15, 5609–5626
22. Emmi, L. *et al.* (2014) Configuring a fleet of ground robots for agricultural tasks. In *Advances in Intelligent Systems and Computing. ROBOT2013: First Iberian Robotics Conference* (Armada, M.A. and et, al., eds), pp. 505–517, Springer
23. Torres-Sánchez, J. *et al.* (2014) Multi-temporal mapping of the vegetation fraction in early-season wheat fields using images from UAV. *Comput. Electron. Agric.* 103, 104–113
24. Torres-Sanchez, J. *et al.* (2013) Configuration and specifications of an unmanned aerial vehicle (UAV) for early site specific weed management. *PLoS ONE* 8, e58210
25. Bauckhage, C. and Kersting, K. (2013) Data mining and pattern recognition in agriculture. *Künstl Intell.* 27, 313–324
26. Yip, K.Y. *et al.* (2013) Machine learning and genome annotation: a match meant to be? *Genome Biol.* 14, 205
27. Sommer, C. and Gerlich, D.W. (2013) Machine learning in cell biology – teaching computers to recognize phenotypes. *J. Cell Sci.* 126, 5529–5539
28. Guyon, I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422
29. Zacharaki, E.I. *et al.* (2009) Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magn. Reson. Med.* 62, 1609–1618
30. Ma, C. *et al.* (2014) Machine learning for big data analytics in plants. *Trends Plant Sci.* 19, 798–808
31. Raza, S-e-A. *et al.* (2015) Automatic detection of diseased tomato plants using thermal and stereo visible light images. *PLoS ONE* 10, e0123262
32. Bergstra, J. *et al.* (2011) Theano: deep learning on gpus with python. *J. Mach. Learn. Res.* 1, 1–48
33. Jia, Y. *et al.* (2014) Caffe: convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678, ACM
34. Kuska, M. *et al.* (2015) Hyperspectral phenotyping on the microscopic scale: towards automated characterization of plant-pathogen interactions. *Plant Methods* 11, 28
35. Peña, J.M. *et al.* (2013) Weed mapping in early-season maize fields using object-based analysis of unmanned aerial vehicle (UAV) images. *PLoS ONE* 8, e77151
36. Wetterich, C.B. *et al.* (2013) A comparative study on application of computer vision and fluorescence imaging spectroscopy for detection of Huanglongbing citrus disease in the USA and Brazil. *J. Spectrosc.* 2013, 1–6
37. Garcia-Ruiz, F. *et al.* (2013) Comparison of two aerial imaging platforms for identification of Huanglongbing-infected citrus trees. *Comput. Electron. Agric.* 91, 106–115

38. Rumpf, T. *et al.* (2010) Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Comput. Electron. Agric.* 74, 91–99
39. Camargo, A. and Smith, J.S. (2009) An image-processing based algorithm to automatically identify plant disease visual symptoms. *Biosyst. Eng.* 102, 9–21
40. Mokhtar, U. *et al.* (2015) Identifying two of tomatoes leaf viruses using support vector machine. In *Information Systems Design and Intelligent Applications* (Mandal, J.K. and et, al., eds), pp. 771–782, Springer India
41. Chen, L. *et al.* (2014) Identification of nitrogen, phosphorus, and potassium deficiencies in rice based on static scanning technology and hierarchical identification method. *PLoS ONE* 9, e113200
42. Huang, K-Y. (2007) Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features. *Comput. Electron. Agric.* 57, 3–11
43. Sankaran, S. *et al.* (2011) Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards. *Comput. Electron. Agric.* 77, 127–134
44. Hillnhütter, C. *et al.* (2011) Remote sensing to detect plant stress induced by *Heterodera schachtii* and *Rhizoctonia solani* in sugar beet fields. *Field Crops Res.* 122, 70–77
45. Casanova, J.J. *et al.* (2014) Development of a wireless computer vision instrument to detect biotic stress in wheat. *Sensors (Basel)* 14, 17753–17769
46. Kersting, K. *et al.* (2012) Pre-symptomatic prediction of plant drought stress using dirichlet–aggregation regression on hyperspectral images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pp. 302–308, AAAI Publications
47. Bauer, S. *et al.* (2011) The potential of automatic methods of classification to identify leaf diseases from multispectral images. *Precision Agric.* 12, 361–377
48. Baranowski, P. *et al.* (2015) Hyperspectral and thermal imaging of oilseed rape (*Brassica napus*) response to fungal species of the genus *Alternaria*. *PLoS ONE* 10, e0122913
49. Schikora, M. *et al.* (2012) An image classification approach to analyze the suppression of plant immunity by the human pathogen *Salmonella* Typhimurium. *BMC Bioinform.* 13, 171
50. Schikora, M. *et al.* (2010) Probabilistic classification of disease symptoms caused by *Salmonella* on Arabidopsis plants. *GI Jahrestagung* 2, 874–879
51. Larios, N. *et al.* (2010) Haar random forest features and SVM spatial matching kernel for stonefly species identification. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, pp. 2624–2627, IEEE Computer Society
52. Ataş, M. *et al.* (2012) A new approach to aflatoxin detection in chili pepper by machine vision. *Comput. Electron. Agric.* 87, 129–141
53. Kruse, O.M.O. *et al.* (2014) Pixel classification methods for identifying and quantifying leaf surface injury from digital images. *Comput. Electron. Agric.* 108, 155–165
54. Bock, C. *et al.* (2010) Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Crit. Rev. Plant Sci.* 29, 59–107
55. Peterson, R.F. *et al.* (1948) A diagrammatic scale for estimating rust intensity on leaves and stem of cereals. *Can. J. Res.* 26c, 496–500
56. González-Pérez, J.L. *et al.* (2013) Color image segmentation using perceptual spaces through applets for determining and preventing diseases in chili peppers. *Afr. J. Biotechnol.* 12, 679–688
57. Wu, D. and Ma, C. (2006) The support vector machine (SVM) based near-infrared spectrum recognition of leaves infected by the leafminers. In *Innovative Computing Information and Control* (2006), pp. 448–451, IEEE
58. Calderón, R. *et al.* (2015) Early detection andq of verticillium wilt in olive using hyperspectral and thermal imagery over large areas. *Remote Sens.* 7, 5584
59. Kaundal, R. *et al.* (2006) Machine learning techniques in disease forecasting: a case study on rice blast prediction. *BMC Bioinformatics* 7, 1–16
60. Behmann, J. *et al.* (2014) Ordinal classification for efficient plant stress prediction in hyperspectral data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* XL-7, 29–36
61. Fortin, J.G. *et al.* (2011) Site-specific early season potato yield forecast by neural network in Eastern Canada. *Precision Agric.* 12, 905–923
62. Gonzalez-Sanchez, A. *et al.* (2014) Predictive ability of machine learning methods for massive crop yield prediction. *Span. J. Agric. Res.* 12, 313
63. Shekoofa, A. *et al.* (2014) Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. *PLoS ONE* 9, e97288
64. Golzarian, M. *et al.* (2011) Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods* 7, 2
65. Dube, T. *et al.* (2014) Intra-and-inter species biomass prediction in a plantation forest: testing the utility of high spatial resolution spaceborne multispectral RapidEye sensor and advanced machine learning algorithms. *Sensors* 14, 15348–15370
66. Cai, J. *et al.* (2015) RootGraph: a graphic optimization tool for automated image analysis of plant roots. *J. Exp. Bot.* Published online July 29, 2015. http://dx.doi.org/10.1093/jxb/erv359
67. Iyer-Pascuzzi, A.S. *et al.* (2010) Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems. *Plant Physiol.* 152, 1148–1157
68. Khazaei, H. *et al.* (2013) The FIGS (focused identification of germplasm strategy) approach identifies traits related to drought adaptation in *Vicia faba* genetic resources. *PLoS ONE* 8, e63107
69. Pieruschka, R. and Poorter, H. (2012) Phenotyping plants: genes, phenes and machines. *Funct. Plant Biol.* 39, 813–820
70. Chéné, Y. *et al.* (2012) On the use of depth camera for 3D phenotyping of entire plants. *Comput. Electron. Agric.* 82, 122–127
71. Raza, S.E. *et al.* (2014) Automatic detection of regions in spinach canopies responding to soil moisture deficit using combined visible and thermal imagery. *PLoS ONE* 9, e97612
72. Römer, C. *et al.* (2012) Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis. *Funct. Plant Biol.* 39, 878
73. Chen, D. *et al.* (2014) Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. *Plant Cell* 26, 4636–4655
74. Cobb, J.N. *et al.* (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. *Theor. Appl. Genet.* 126, 867–887
75. Honsdorf, N. *et al.* (2014) High-throughput phenotyping to detect drought tolerance QTL in wild barley introgression lines. *PLoS ONE* 9, e97047
76. Topp, C.N. *et al.* (2013) 3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1695–E1704
77. Rasheed, A. *et al.* (2014) Genome-wide association for grain morphology in synthetic hexaploid wheats using digital imaging analysis. *BMC Plant Biol.* 14, 128
78. Yang, W. *et al.* (2014) Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5, 5087
79. Guo, W. *et al.* (2015) Automated characterization of flowering dynamics in rice using field-acquired time-series RGB images. *Plant Methods* 11, 7
80. Suárez, L. *et al.* (2010) Detecting water stress effects on fruit quality in orchards with time-series PRI airborne imagery. *Remote Sens. Environ.* 114, 286–298