Feature-Ensemble Based Novelty Detection for Analysing Plant Hyperspectral Datasets

DOI:

[10.1109/JSTARS.2017.2788426](https://doi.org/10.1109/JSTARS.2017.2788426)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

AlSuwaidi, A., Grieve, B., & Yin, H. (2018). Feature-Ensemble Based Novelty Detection for Analysing Plant Hyperspectral Datasets. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(4), 1. [99]. <https://doi.org/10.1109/JSTARS.2017.2788426>

Published in:

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Feature-Ensemble-Based Novelty Detection for Analyzing Plant Hyperspectral Datasets

Ali AlSuwaidi, *Student Member, IEEE*, Bruce Grieve, and Hujun Yin, *Senior Member, IEEE*

Abstract—Recently, there has been a significant increase in the use of proximal or remote hyperspectral imaging systems to study plant properties, types, and conditions. Numerous financial and environmental benefits of using such systems have been the driving force behind this growth. This paper is concerned with the analysis of hyperspectral data for detecting plant diseases and stress conditions and classifying crop types by means of advanced machine learning techniques. Main contribution of the work lies in the use of an innovative classification framework for the analysis, in which adaptive feature selection, novelty detection, and ensemble learning are integrated. Three hyperspectral datasets and a nonimaging hyperspectral dataset were used in the evaluation of the proposed framework. Experimental results show significant improvements achieved by the proposed method compared to the use of empirical spectral indices and existing classification methods.

Index Terms—Ensemble learning, feature selection, hyperspectral imaging (HSI), novelty detection (ND), plant monitoring, remote sensing, support vector machine (SVM).

I. INTRODUCTION

THE discovery of imaging spectrometry has triggered a great deal of scientific effort, focusing on the use of spectral information [1] in addition to spatial data. Hyperspectral imaging (HSI), a branch of multivariate imaging [2], employs spectroscopy and remote imaging technologies to capture the optical properties of the target with multiple spectral representations [3]. Interest in this emerging area has increased in research institutions and industries, due to the advantages gained from sensing a large number of narrow spectral bands and a wider range of the electromagnetic spectrum. Moreover, HSI has been exploited in increasing number of applications from remote and proximal sensing [4]–[7], chemical processes [8], medical imaging [9], and industrial processes [10] to agricultural and environmental monitoring [6], [7], [11]. It is worth noting that hyperspectral images can be acquired using four different configurations [2], [12]: single shot, area, line, and point scanning.

Activities in HSI have increased recently due to numerous environmental and financial advantages it can bring [6], [7], [13].

Manuscript received May 31, 2017; revised September 29, 2017 and December 19, 2017; accepted December 20, 2017. (*Corresponding author: Hujun Yin.*)

The authors are with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: ali.bghalsuwaidi@postgrad.manchester.ac.uk; bruce.grieve@manchester.ac.uk; hujun.yin@manchester.ac.uk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JSTARS.2017.2788426

For examples, proximal HSI systems (i.e., ground-based) have been used in agriculture to study plant properties and conditions for the purpose of monitoring plant health, detecting diseases, preventing from spreading, and precision control of a herbicide process (i.e., spraying weed species only, hence leading to reduction of herbicide amount, weed species' resistance to herbicide, and pollution effect of herbicide to the environment). Remote HSI systems in unmanned vehicles and satellites have been utilized in several applications such as urban planning, disaster management, and change detection because such systems can observe large areas as well as can access harsh environment compared to proximal HSI systems [14], [15].

Recently, researchers have shown an increased interest in developing efficient and effective analysis tools for hyperspectral images or data, since a large amount of HSI data is being generated and it is difficult, if not impossible, to analyze the information directly from the pixel values [16], [17]. Increasingly machine learning techniques are applied due to the scale and complexity of the problems. It is true to say that the more data collected, the greater the regions of interest it may cover; however, not all collected data and/or variables necessarily contain useful information relating to the problem being examined. A number of methods have been developed for dimensionality reduction and are applicable for hyperspectral data [18]–[20]. One of them involves extracting the most differentiable information to the problem from the data. In this regard, the most significant information is statistically selected after transforming the original data into another feature space, hence reducing the dimensionality. Principle component analysis (PCA) and Fisher's discriminant analysis are classic examples. The dimensionality problem can also be alleviated by using feature selection [21], [22]; a process of selecting a relevant subset of features (e.g., wavelengths in HSI data) and removing irrelevant and/or redundant ones based on certain evaluation criteria.

According to the previous studies [23]–[25], feature selection can improve the performance of classification, compared with using the entire feature space, in addition to reducing computational costs. However, classification performance and required number of features vary among individual algorithms to select and discard redundant or irrelevant features. This means that there is no single, standard method to define the best algorithm for a specific set. The lack of robustness in this regard can be solved by combining several feature selection algorithms. The process of combining several learning systems in machine learning is called ensemble learning [26]–[28]. The combination can be performed either in weighted or nonweighted sense, in order

to obtain more robust classification performances compared to using single learning systems.

In recent years, there has also been growing interest in fusing several properties (e.g., texture and spectral information) of single data source (termed feature fusion) or certain properties of multidata sources (SAR and optical sensors) under centralized decision (termed decision fusion) to improve classification accuracy [29]–[31]. Feature fusion schemes retain the best features to improve the classification accuracy locally, while decision fusion schemes integrate local classifiers to improve the overall performance. Several studies have been published in this regard [29]–[31]. In [29], the dimensions of five different classes were reduced and then used to generate the final decision through majority voting and a neural network. Feature ranking and selection were used in [30] to identify the optimal subset of features of each classifier and then the final decision of all classifiers was combined using a sum rule. The study [31] proposed an adaptive differential evolution decision to combine different features extracted from hyperspectral and LiDAR datasets.

A large and growing body of literature has investigated different approaches for classifying remotely sensed imagery, including supervised (label information available), unsupervised (no label information available), and semisupervised classification approaches [32]. In the supervised approach, a support vector machine (SVM) is widely used due to its good generalization ability. However, the SVM is developed primarily for binary cases, while the remotely sensed data often involve multiple classes. Methods to handle multiple classes have been proposed in the literature by using multiple SVMs to overcome this issue, in either parallel format (one against all or one against another) or hierarchical tree based (e.g., balanced branches and one against all) [33]. The unsupervised approach relies on clustering training samples to obtain classification map. Typical methods include K-means, neural networks, fuzzy C-means [32], and swarm-based fuzzy C-means [34]. Attention on the semisupervised approach has increased due to wide availability of unlabeled data in addition to limited labeled training samples [35]. The use of unlabeled samples can help achieve more accurate modeling of class distributions and thus improving the classification performance [36].

Novelty detection (ND) [37]–[41], also termed one-class classifier, is a machine learning approach to detect abnormality where only normal samples are available and used to construct the classifier or prediction model. The main goal of ND is to recognize how a test sample is deviated from the training ones, that is, how abnormal is the test sample compared to the normal training samples. ND has been used in monitoring high integrity systems such as jet engines, whereas distributions of abnormal classes are difficult if not impossible to obtain [41].

This work focuses on analyzing and classifying hyperspectral datasets, captured by several sensors (proximal sensing and remote sensing HSI systems, and nonimaging spectroradiometer) with a proposed framework, and comparing the results with that of using existing empirical indices and other classification methods. Advanced machine learning techniques such as feature selection, ND, and ensemble learning are employed in this framework to demonstrate how combining machine

learning techniques can be used for classifying hyperspectral data. Main contribution lies in combining a feature-ensemble method proposed in our previous work [17] and an ND method based on the SVM. An ND by SVM (ND-SVM) (calibrated form of the one-class SVM) classifier is used to construct the prediction model instead of conventional multiclass classifiers in order to avoid the limitation of the latter in unbalanced data cases, since here the applications are to differentiate control and diseased or stressed plants. The advantages have been verified with markedly improved discrimination achieved.

The remainder of the paper is organized as follows. Background and related work are reviewed in Section II. Section III describes the machine learning techniques relevant to this work. Section IV presents the imaging and nonimaging systems used, datasets, and the proposed classification framework. The experimental results and discussions are given in Sections V and VI, respectively, followed by the concluding remarks in Section VII.

II. BACKGROUND AND RELATED WORK

Remote sensing can be defined as “the art of science of telling something about an object without touching it” [42] or “the observation of a target by a device separated from it by some distance” [11]. It can be classified as passive or active based on the sensor; the former gathers the radiation that is incident on the sensor used (i.e., no radiation is transmitted), while the latter gathers the reflected radiation to the sensor after transmitting it to the targeted object [4], [5]. Airborne visible–infrared imaging spectrometer (AVIRIS) [43] is an example of the passive sensor, while TerraSAR-X SAR [44] is an example of the active sensor. Typically, aircrafts or satellites are used to gather remotely sensed data and advantages include, amongst others, covering large areas in single imaging and the applicability over a wide range of applications such as change detection [15], environmental monitoring [45], and crop classification [46], [47].

On the other hand, proximal sensing involves collecting information of the object of interest in a nondestructive manner at close range [48]. Advantages include capturing detailed information especially in terms of spatial resolution and generating replicas of the captured data in a short time and at a low cost. In the last two decades, a great number of studies have been conducted using such systems to study plant properties and conditions. Several spectral indices have been proposed in this regard. The primary formulations are the simple ratio (SR) between reference and index reflectance and the normalized difference ($N_{\text{orm}}D_{\text{iff}}$) between them [49]:

$$\text{SR} = \frac{R_{\text{ref}}}{R_{\text{index}}} \quad (1)$$

$$N_{\text{orm}}D_{\text{iff}} = \frac{R_{\text{ref}} - R_{\text{index}}}{R_{\text{ref}} + R_{\text{index}}} \quad (2)$$

where R_{ref} and R_{index} represent the reference and index reflectances at a specific wavelength, respectively. There are a few indices that are adapted from these primary forms, such as soil-adjusted vegetation index for studying certain properties [50].

Such existing indices have been used in many applications such as detection of plant stress [49], [51], [52], weed management [6], [7], [13], and diseases identification and detection [53]. Studies of plant stress detection were carried out by Gitelson and Merzlyak [51], and Sims and Gamon [49]. In the first study, a positive correlation in the visible spectrum was found between leaf pigment contents and seasonal stress levels; according to this study, the pigment concentration levels of stressed Chestnut and Maple leaves in seasonal conditions only reached 40–50% of that under the relaxed condition. This means that the variation in the spectral profiles, due to pigment variation, can be used to detect and identify the stressed leaves. The second study focused more on green pigment levels (i.e., chlorophyll) as they provide insight into leaf condition. It has been reported that leaf area and structure play an important role in estimating green pigment. Several indices were developed in the past to measure different leaf pigments. However, no clear evidence was found for any relationship between a leaf pigment and its structure, since most of these indices were tested for a limited number of species (one or few). The corrected indices that were used in the second study showed better estimation of green pigment over a wide range of leaves, thus detecting stress with no extensive calibration. In addition, indices were introduced to detect water stress level such as water index (WI) [54] and normalized difference WI (NDWI) [55].

Previous studies on weed management have revealed the importance of using proximal hyperspectral sensors in agricultural applications [6], [7], [13]. Reduction in the workforce needed and the amount of herbicide used and improvement in the production process are notable advantages of weed management. Many studies have utilized spectral indices to differentiate between weeds and other crops. For instance, the normalized difference vegetation index (NDVI) was used along with the hue of the entire spectrum, the difference between near infrared (NIR) and red spectra, and the difference among NIR, red, and hue, to produce an automatic proximal weed identification system. This system identifies weed locations in order to control the herbicide-spraying process.

In terms of disease identification and detection, several indices have been proposed and used. On the one hand, a combination of several spectral indices could be used not only to differentiate between healthy and unhealthy leaves, but also to determine disease severity level [56], [57]. Usually, a difference in spectral profiles between infected and healthy leaves can be detected over different regions and the difference in spectral index values can lead to identifying healthy from unhealthy leaves. On the other hand, specific disease indices could also be used to differentiate the health condition of the leaves. This is exemplified in a study undertaken by Mahlein *et al.* [53], where four different indices were proposed [i.e., healthy index (HI), sugar beet rust index, Cercospora leaf spot index, and powdery mildew index] to differentiate between healthy and unhealthy sugar leaves. It should be mentioned that a combination of single wavelength and the normalized difference of two wavelengths was considered in this study. Moreover, a distance criterion based feature selection algorithm (Relief-F) was used to find the best combination for discrimination.

In addition to these indices, machine learning has been also employed. For instance, an advanced machine learning was employed for disease detection and identification [56]. A multi-class SVM (LIBSVM) with radial basis function (RBF) kernel was used to construct a classification model that can detect and identify healthy sugar leaves, Cercospora leaf spot, powdery mildew, and sugar beet rust. It should be noted that several spectral vegetation indices (with different physiological parameters) were calculated and used as features instead of the spectral responses. Moreover, a parallel method (one against all strategy) was used to extend the binary SVM to a multiclass SVM. The overall classification accuracy suffered due to difficulties in class separability. AlSuwaidi *et al.* [16] used feature selection and multiclass SVM with RBF kernel for plant speciation. Feature selection reduced the feature space into optimal subset of features and helped the multiclass SVM to improve the classification accuracy, compared to the use of all wavelengths as well as spectral vegetation indices.

It should be noted that most of these indices, if not all, have been used in the literature to analyze remotely sensed images. For example, NDVI is widely used in satellite remote sensing to describe the density and the health of vegetation. Usually, NDVI values range from -1 to 1 , where low values (i.e., 0.1 or less) indicate nonvegetated areas, moderate values (i.e., 0.2 – 0.5) sparse vegetation, and high values (i.e., > 0.6) dense vegetation. Another example is the use of NDWI to detect water stress from space [55], using both visible and NIR regions. Although NDWI is less sensitive to atmospheric scattering compared to NDVI, it is not a substitute (i.e., complement) to the latter index. Moreover, a previous study was conducted using water indices to estimate the hazards of wildfire from remotely sensed data [58]. The authors suggested that using such technique with accurate estimation would help reduce hazardous effects on economy, environment, and social life. In addition, earth observation-1 (EO-1) Hyperion imagery was used to detect orange rust over Mackay, Qld., Australia [57]. Several narrow-band indices, existing and newly proposed by the authors, were tested in this study in order to select the optimal ones with high separability between the infected and noninfected orange rust areas.

III. MACHINE LEARNING TECHNIQUES

This section outlines machine learning techniques relevant to this work. It is divided into three sections: starting with a description of a feature selection process, followed by an overview of ensemble learning, and an introduction of classifiers (i.e., conventional and one-class SVMs).

A. Feature Selection

The need for efficient and effective analysis methods has been elevated along the advances in the HSI system as large amount of data is being generated and it is difficult to analyze the information directly from pixel values. This problem can be addressed by using feature extraction or selection techniques. Feature selection can be defined as a process of selecting only a subset or relevant features by removing irrelevant and redundant ones [21], [22]. It aims to identify a minimal subset of available fea-

tures that give improved or similar (if not better) performance as using the full set of features based on certain evaluation criteria. A large and growing body of literature has investigated the process of selecting relevant information. As a result, several algorithms have been introduced [23]–[25]. There are numerous studies [21], [22], [59], [60] that suggest how relevant features x_i can be selected from a feature space $\mathbf{x} = \{x_1, \dots, x_n\}$ for a specific dataset $\mathbf{X}$. For example, the k th feature (x_k) can be considered as relevant if it can be used to distinguish between two specified different classes $Y = \{y_1, y_2\}$ in the dataset. The main advantages of using only the relevant subset of all features include reduction in processing time as well as improving prediction or classification performance [21].

There are a large number of published studies describing feature selection algorithms; however, all of these studies have the same principle in the process, which can be described in four steps: search organization, subset evaluation, stopping criteria, and result validation [21], [22], [60]. The first step determines whence the search is initiated (e.g., growing an empty subset or shrinking a full one) in the feature space, as well as how the feature space is explored (e.g., complete, random, or heuristic) to generate different subsets. The second step involves using dependent and independent criteria to evaluate the goodness of the generated subsets, in order to estimate the optimal subset. In the latter criteria, data characteristics such as distance, information, or consistency are used for the evaluation, while error probability or accuracy is used in the former. The third step determines when the process should be halted, by reaching a certain threshold, or when growing or shrinking a feature subset will not further improve the prediction performance, or in completing the search procedure. The last step is usually used to improve the total process of selection if possible, i.e., checking whether the selected subset, or the evaluation criterion, or the stopping criterion is optimal.

Feature selection can be generally categorized, based on evaluation criteria, into filter, wrapper, and embedded models [21], [22], [59]–[62]. All these models tend to maximize their evaluation criteria $J(\cdot)$ in order to define a significant subset of features S_o :

$$J(S_o) \geq J(S_j) \quad \forall S_o, S_j \subseteq \mathbf{X}; j = 1, 2, \dots, m \quad (3)$$

where S_j represents the j th subset of m total subsets S generated in the feature selection process.

In the filter models, the evaluation criterion is solely dependent on data characteristics, that is, rank features based on their content and then select the best among them without using any mining or classification algorithms. Evaluation criteria in the filter model can be listed as distance, information, dependence, and consistency. Furthermore, the output of this model is given as either weighted features or an optimal feature subset. Wrapper models evaluate the goodness of features using mining or classification algorithms to determine the optimal features subset. The selected features (iteratively) are tailored for a specific mining algorithm; this means that the selected features may not perform well with other mining algorithms. Embedded models combine the both to take the advantages of the filter and wrapper models. They can avoid predefined stopping criteria and can

TABLE I
GENERALIZED FEATURE SELECTION ALGORITHM

Generalized feature selection algorithm	
Requirements:	
$\mathbf{X}$:	dataset
$\mathbf{x}$:	feature space with $x_1, \dots, x_n$ features
Y :	labels
J :	evaluation criteria
CA :	classification algorithm (Wrapper and Embedded)
Feature Selection Process:	
S_{initial} :	initialize feature subset (empty or full)
Start:	
S_j :	add or remove feature to the subset based on S_{initial}
Evaluate($S_j, \mathbf{X}, J$)	
Evaluate($S_j, \mathbf{X}, CA$)	
If $J(S_j)$ is better than $J(S_{\text{initial}})$	
If $CA(S_j)$ performance is better than $CA(S_{\text{initial}})$	
$S_o = S_j$:	optimal subset
Repeat if stopping criteria = false	
end	

handle a large amount of data. Moreover, embedded models use data characteristics to generate several feature subsets with high potential and utilize mining algorithms to select the most superior subsets. Filter models can produce acceptable results in less time, compared to the other models. Pearson's correlation [59], information gain [21], [22], correlation feature selection (CFS) [25], relief and its extension [23], [24], and sequential forward generation [59] are examples of feature selection algorithms. The pseudocode of a generalized feature selection process is presented in Table I.

It is now well established from previous studies (e.g., [16], [25]) that feature selection has positive effects on classification or prediction performance as well as processing time. However, most of these studies consider single data source/properties problem only. In a multisource/property problem, a proper fusion method is needed in order to combine extracted features from each data source since the collected data from these sources may differ in terms of spectral, texture, temporal, and so on [63]. Feature concatenation is the simplest method to combine these extracted features but the performance is not guaranteed (i.e., it may worsen the performance since it does not consider the structure differences of the feature space) [64]. Several methods exist for fusing features of multiple data sources or properties. For example, a feature fusion based marginalized kernel (i.e., softmax regression based) has been proposed to combine different properties of QuickBird imagery [65]. Spectral, texture, and local-self similarity descriptors from each object are used to encode feature vector (as bag-of-visual words) and then estimate object-to-class and class-to-class similarities for feature fusion purpose. Generalized graph-based fusion technique is another paradigm of a feature fusion scheme [64]. A kernel PCA (KPCA) was used in this study to normalize the morphological features extracted from HSI images, spatial and elevation (LiDAR) information individually, thus reducing their dimensions and noise levels for final fusion. Another example of feature fusion is presented in [63], where decision tree was used to fuse the outputs of the extracted medium-resolution SAR and

optical subpixel features in order to map forest/land coverage and to detect deforestation or forest degradation.

B. Ensemble Learning

Ensemble learning has been studied for more than 20 years, with the primary goal of improving the prediction performance [26]–[28], [66]. It can be defined as a process of combining various learning algorithms, either weighted or unweighted, to enhance the prediction performance over single learning systems. The main advantage of such an approach is that the risk of selecting a weak or least performing learner is minimized. The diversity of the learning algorithms that generate uncorrelated error patterns needs to be promoted in the ensemble learning in order to improve the performance [26], [28]. The diversity can be ensured using either resampling techniques or different features. Ensemble diversity can be measured by several quantitative methods [28], [66]. Correlation and Q -statistics are two examples to measure the diversity through the probabilities:

$$r = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(FN + TN)(TP + FN)(FP + TN)}} \quad (4)$$

$$Q = \frac{(TP \times TN) - (FP \times FN)}{(TP \times TN) + (FP \times FN)} \quad (5)$$

where TP and TN represent the true positive and true negative rates, i.e., ratios of the samples that are correctly and wrongly classified by the learning algorithms to the total number of samples, respectively. FP, FN (i.e., false positive and false negative) represent the ratios of the samples that are correctly classified by at least one learning algorithm to the entire positive and negative samples, respectively, and the summation of all ratios is unity.

Various methods can be used to construct ensembles [28]. Some of them are general, using any learning algorithm, while others are limited to particular learning algorithms. Bagging is one of the simplest ensembles to construct, where different input subsets are randomly replicated from the original input dataset with the goal of increasing diversity. Boosting is another general ensemble that boosts the performance of weak learning algorithms. Stacking is an old method and thus less preferable to bagging and boosting, as there is no standard procedure to implement it. This method is divided into two stages (i.e., ensemble and learning) and the final result is achieved by combining the entire outputs of each learner.

Different theories exist in the literature regarding how the eventual outputs can be combined (i.e., “decision fusion”). On the one hand, the simplest scheme is to use the hard fusion [27], [28]. Hard fusion includes weighted or unweighted versions of majority voting for class labels case and weighted or unweighted versions of the mean in continuous output case (or regression). On the other hand, soft decision schemes, such as linear opinion pool and logarithmic opinion pool, use the posterior probabilities of each classifier (without relaying on class label of each of them) to generate the central decision (final ensemble result) [67]. In a recent study, a hierarchical decision fusion scheme is proposed to classify an outdoor scene [68]. This scheme consists of two stages: the first uses Gabor feature extraction to extract the scenes into artificial or natural images,

while the second stage fuses the decisions of gradient binary patterns, grey level co-occurrence matrix, and color features for further classification (i.e., to identify forest, coast, mountain, and open country scenes).

C. Novelty Detection

The role of one-class classification in pattern recognition has received attention across a number of disciplines in recent years. One-class classification, commonly used for ND, can be defined as the process of modeling a well-distributed class in order to classify an unseen testing class. It is considered as a special case of a classification problem with the availability of one known class instead of two or more classes to train the model (i.e., classifier). In ND, the prediction performance is measured based on the ability of the constructed model to recognize samples of the normal class, i.e., samples characteristics similar to that of those used in training, from the other samples, called abnormal or novel, or outliers [37]–[41]. It is worth knowing that the constructed model reflects the coverage of the normal class in the feature space [69]. In contrast, the abnormal class is totally absent (sparsely distributed in some cases), and hence, it is very difficult to obtain these samples, if not impossible, such as in high integrity systems. The abnormality of such systems would have to be obtained by damaging the systems in many possible ways, and that is impractical and can lead to huge losses.

ND techniques can be classified into five categories: density-based, distance-based, reconstruction-based, domain-based, and information-based [38]. The first category uses density estimation to decide whether an unseen test sample is considered as normal or abnormal; the area with high density indicates high probability of sample to be normal. Gaussian mixture models and kernel density estimators are two popular examples of this category. Test samples are classified based on the estimated parameters in the former, for example, using the maximum likelihood from a mixture of Gaussian distributions [70]. On the other hand, Parzen windows centered on each training point with the same variance are often used to estimate the probability density function in the latter case [71]. The second category uses the distance between the test and training samples to identify if the test samples fall under normal or abnormal categories; abnormal samples fall far away from normal samples. K nearest neighbor is a common example of a distance-based method where different distance measures can be used such as Euclidean and Mahalanobis [72].

The third category defines the abnormality based on the difference between the actual values of a model and observations, classifying as normal if not exceeding certain threshold, otherwise abnormal. One way to implement ND for the reconstruction category is by transforming the data into low dimensional space where a small number of features are used to describe and model the normal samples. In the classifying stage, the correlation between a testing sample and the parameters of the trained model are used to identify the class type with a large value representing the abnormal class. The fourth category uses the training samples to construct the model, thus generating the normal class boundaries, also known as the class domain. Everything lying

within the generated boundaries is classified as a normal sample, while anything outside is classified as abnormal. The SVM is a popular paradigm for this category where it is used to maximize the margin from the origin [73], [74], or surrounding the normal data with a spherical boundary [69]. The last category uses the information content of the training set as an indicator to recognize the normal samples from the abnormal one. That said, the presence of abnormal samples will lead to significant change in the information content. Shannon's entropy [75] is usually used to differentiate between the normal and abnormal samples. The test samples will be considered as abnormal if their removal will lead to a reduction in the entropy value.

ND has been employed in several domains such as industrial applications, medical monitoring, and remote sensing [38]–[41]. In industrial applications, ND has been used in monitoring and analyzing high interiority systems such as jet engine where the normal samples are most dominant [39], [41]. The presence of failure samples is rare if not completely absent in this case; thus, it is difficult to use conventional, two-class classifiers. ND is the best choice in this case since a robust model, based on normal operation samples only, can be implemented to assess the health of the system and identify any strange activities in advance. In medical applications, ND has been used to detect any unusual record and raise a flag of potential health problems [38]. Remote sensing is another domain where ND can be employed, more specifically in change detections [40]. Usually, the change characteristics such as new buildings, new desertification areas, and forest fires are often unknown and only the characteristics of unchanged areas are available. In this case, the unchanged areas can be considered as normal samples and used in the training to identify any unexpected changes.

In the proposed framework, a domain-based ND-SVM is used. The SVM is considered as one of the best classifiers [76] and can deal with dimensionality problem effectively, hence reducing the risk of overfitting [77]. In a binary problem, the SVM is used to generate the best hyperplane that maximizes the margins. In addition, the kernel trick can be used to deal with the nonlinear cases in order to find the best hyperplane. Several kernels are available such as RBF and polynomial:

$$\text{RBF} : K(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|^2}, \gamma = \frac{1}{2\sigma^2}$$

$$\text{Polynomial} : K(\mathbf{x}, \mathbf{y}) = ((\mathbf{x}, \mathbf{y}) + 1)^n, n \in \mathbb{Z}^+. \quad (6)$$

However, in ND, only one class is available; this case is highly unbalanced, so modifications have to be made to the conventional SVM in order to deal with unbalanced data. The approach considers a one-class SVM introduced in [73] and [74], where the output value for a testing sample is 1 if it is within the boundary of the normal class and -1 otherwise. Such a model can be generated by solving the following quadratic problem with regard to training set $\mathbf{X}_{\text{train}} = \{\mathbf{x}_1, \dots, \mathbf{x}_t\}$:

$$\begin{aligned} \min_{\omega \in \phi(\mathbf{X}_{\text{train}}), \xi \in \mathbb{R}^t, b \in \mathbb{R}} \quad & \frac{1}{2} \|\omega\|^2 + \frac{1}{\vartheta_t} \sum_i \xi_i - b \\ \text{subject to : } & (\omega \cdot \phi(\mathbf{x}_i)) \geq b - \xi_i \quad \forall \xi_i \geq 0 \end{aligned} \quad (7)$$

where $\omega, b, \phi(\cdot), \xi$ represent weighting vector, bias, mapping function, and nonzero slack variable, respectively. Moreover, $\vartheta \in \{0, 1\}$ represents the lower bound of the support vectors as well as the upper bound to the outliers that used to determine the flexibility of the decision boundary. In this case, the decision function $f(\mathbf{x})$ can be solved using the weight vector as well as the bias:

$$f(\mathbf{x}) = (\omega \cdot \phi(\mathbf{x}_i) - b). \quad (8)$$

The value of decision function $f(\mathbf{x}) \in \{\pm 1\}$, where 1 represents the normal class and -1 the abnormal class. It should be mentioned that the false positive and negative errors have to be reduced in order to obtain a good novelty model. Constructing a dual optimization problem with the use of appropriate kernels will allow to consider the training samples located in nonlinear space [37]. In this case, the decision function can be written as

$$f(\mathbf{x}) = \left(\sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) - b \right) \quad (9)$$

where α_i represents the Lagrangian multiplier and $\mathbf{x}_i$ represents the support vector. A quadratic problem routine can be used to solve the dual problem, thus finding the parameters that determine the best separation boundary:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{ij} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to : } & \frac{1}{\vartheta_t} \geq \alpha_i \geq 0, \sum_i \alpha_i = 1. \end{aligned} \quad (10)$$

The one-class SVM [73] was proposed as an extension of the binary classifier for ND problems, in which there are only samples for one class. The ND-SVM is a particular one-class SVM with RBF kernel and calibrated output [39]. The key difference between the ND-SVM and the one-class SVM is that the value of the decision function is further calibrated to class probability using isotonic (i.e., nondecreasing) regression in the ND-SVM [39]. Isotonic regression is a nonparametric form of regression and is an intermediary approach between binning and sigmoid fitting. The calibrated output in the ND-SVM case is achieved by employing a pair-adjacent violator routine, where the isotonic regression is performed on the sorted training samples (based on their scores) to find the stepwise-constant isotonic function (mean-squared error based). The calibrated output, mapped onto the range $[0, 1]$, is then used to define the threshold as well as to determine the best parameters of our classifier.

IV. MATERIALS AND THE PROPOSED METHOD

This section describes the materials and the proposed methods. It first emphasizes on specifications of the HSI systems used to capture the hyperspectral datasets and then describes the datasets used for the evaluation purpose, followed by the description of the proposed classification framework.

A. Nonimaging HSI

A hand-held spectro-radiometer (ASD FieldSpec Pro FR spectrometer provided by the Analytical Spectral Devices Inc.)

was used to extract leaves spectral profiles [53], [78]. This device operates over a spectral range of 350–1100 nm with spectral resolution of 3 nm; however, only the spectral range of 400–1050 nm was used in the analysis as the extremes suffer from noise. Moreover, this device needs to be calibrated before acquiring the spectral profiles in order to extract them well. Calibration involves warming up the device for 90 min to improve the quality, then calibrating dark noise using a barium sulfate surface. No preprocessing was performed, since the device was operated in a controlled environment (i.e., constant light and temperature conditions).

B. Imaging HSI

Hyperspectral images were collected from three HSI systems: University of Manchester (UoM) system, University of Bonn (Bonn) system, and AVIRIS sensor. The UoM is described in [79] and [80]. This system consists of a monochromatic digital camera with Peltier cooling and provides images with a spatial resolution of 1024×1344 pixels. Moreover, a fast liquid crystal tunable filter together with an infrared blocking filter were mounted in front of the lens to control the spectral transmission electronically and to prevent leakage. The intensity response of each pixel was recorded using a 12-bit analog-to-digital converter. This system was configured to capture 33 images over wavelength ranging from 400 to 720 nm in 10 nm steps (peak transmission) and the bandwidth (FWHM) varies over the wavelength range. It should be noted that this system operates in a controlled environment (dark room), in order to minimize the effect of unwanted noises.

The Bonn system is a line scanning system that generates a maximum effective 1600 pixels and it is surrounded by six lamps (provided by the Analytical Spectral Devices Inc.) to provide homogenous illumination [81]. This system operates over the range of 400–1000 nm with a 2.8 nm spectral resolution and a spatial resolution of 0.29 mm. It is worth knowing that the data were captured in controlled environment (i.e., dark chamber) in order to reduce noise and realize optimal reproducible illumination. More information about the orientation and preprocessing can be found in [53] and [81].

The AVIRIS sensor, a first operational hyperspectral instrument, was developed, owned, and operated by NASA JPL. It is a whiskbroom sensor that uses line arrays of detectors to capture an image with 12 km swath width at 224 spectral bands ranging from 360 to 2500 nm with an approximate spectral resolution of 10 nm [43]. This sensor has been flown on four different aircraft platforms at different altitudes resulting in images with different spatial resolutions (i.e., approximately 2–4 m at low altitude and 20 m at high altitude). AVIRIS data are stored in 16-bit integers and the data products include level 1B (i.e., orthorectified) and level 2 (i.e., atmospherically corrected). This sensor has been utilized in a number of applications, for instances, mineral exploration, agriculture, atmospheric analysis, Geology, and sensors calibration.

C. Datasets

Four HSI datasets were considered in our experiment: a nonimaging and three imaging datasets. Regarding the nonimaging

Fig. 1. Leaf samples. RGB representation of UoM hyperspectral dataset (*Arabidopsis*) captured by the UoM HSI system in January 2017 for the illustration purpose. The leaves were placed flattened to maximize the reflectance.

dataset, the spectral reflectance of several sugar leaves (controlled and diseased; *Cercospora*, rust, and powdery mildew) was measured using a hand-held spectro-radiometer [78]. The spectral profiles of controlled and diseased sugar leaves were recorded from day 1 to day 21 with different severity levels and normalized using a white reference surface (i.e., ratio of spectral profile of sample to spectral profile of white surface). Total of 2516 samples were extracted using this device, 630 healthy samples (i.e., considered as normal) and 1886 unhealthy samples (i.e., considered as abnormal). Moreover, 60% of the normal sample was used for training, and the remaining 40% of the normal and all the abnormal samples were used for testing and validation. For simplicity, we called this dataset Bonn-SPEC; more information on how the data was recorded can be found in [78] and [53].

For the UoM dataset, three different images were captured; scene (actual image), dark noise (electronic noise), and flat field (illumination distribution) images. The scene image consisted of six leaf samples placed flatten on the sample plate; two of which were of normal conditions while the other four were not (two heat and two cold stressed). The plate also included a reference color patch for further calibration. It should be mentioned that all images were prevented from saturation (dynamic range management). Moreover, the scene image was spectrally normalized using flat field and dark noise images to enhance the quality of the image. In the example of a scene image, shown in Fig. 1, leaves in the left (top and bottom) represent the controlled leaves, while leaves in the middle and right (top and bottom) represent cold and heat stressed, respectively. In total, 18 replicated scene images were obtained on the same day. A total of 648 samples were extracted from all these scene images and divided into two groups (i.e., normal and abnormal). The normal group was represented by 216 controlled samples, while the abnormal group by 432 stressed samples (216 cold and 216 heat stress samples). Only green areas were considered in extracting these samples. In addition, 60% of the normal group was used for training, and the remaining 40% of the normal and all the abnormal groups were used for testing and validation.

Second imaging dataset (Bonn) contains of healthy and unhealthy (*Cercospora* and rust) sugar leaves captured in five different dates (i.e., days 7, 9, 11, 13, and 15) [82]. Similar to the UoM dataset, three different images were captured for

Fig. 2. RGB representation of healthy sugar leaves. The data are courtesy of Bonn University, captured in March 2013. The leaves were placed flattened to minimize the reflectance issue.

Fig. 3. Indian Pines. Ground truth reference map.

the normalization purpose and the dataset was prevented from saturation. The scene images were limited to four leaves, which were flattened, as shown in Fig. 2. Each scene was limited to one condition only (either healthy or unhealthy). The abnormal group was represented by 320 samples (combining *Cercospora* and rust samples), while the normal group was represented by 160 samples. Again, 60% of the normal sample was used for training, and the remaining 40% of the normal and all the abnormal samples were used for testing and validation.

Third imaging dataset was captured by the AVIRIS sensor in June 1992 of Indian Pine test site, which has been widely studied in the literature of hyperspectral remote sensing. This scene is a subset of a full scene that covers portions of Northwestern Tippecanoe County, IN, USA. The size of the subset is 145×145 pixels with a spatial resolution of 20 m covering 16 different crops and provided in the ground truth reference map, shown in Fig. 3. Moreover, 20 bands, covering water absorption regions, were removed leaving the remaining bands for analysis and classification. In addition, the classification was performed using one against all strategy; one of the classes was considered as the normal class, while the remaining classes considered as the abnormal class. Similarly, 60% of the normal group was used for training, and the remaining 40% of the normal and all the abnormal groups were used for testing and validation.

D. Proposed Framework

The novelty of our proposed framework lies in utilizing the advanced machine learning techniques to design a robust classification framework for hyperspectral data, presented in

Fig. 4. The employed techniques include feature selection, ND, and ensemble learning. Prior to commencing the classification framework, the need of the scaling process is checked to ensure minimal variation in the dataset. After the scaling process, the input data are divided into N subsets using “jackknife” or k -fold data split [28]. On completion of data splitting, the process of selecting a relevant subset of wavelengths is carried out using several feature selection algorithms (as used in [17]), called feature selection pools, and only the best performing algorithm is retained in each pool. The main goal of using several feature selection algorithms in each pool is to improve the robustness of the selection process in the respective pool since there is no single feature selection algorithm that can provide sufficient result across all pools. Moreover, the splitting and selecting processes are used together to ensure the diversity of the ensemble process. Several articles reported the importance of resampling the input dataset or using different features to promote diversity; however, no one has used the combination of both to the best of the authors’ knowledge. Once the wavelengths are selected, it is passed to the ND classifier for classification. It should be noted that the ND-based SVM, an extended version of the original one-class SVM proposed in [39] (will call it ND-SVM), classifier is used in our framework since we are interested in the description of the unseen data (i.e., healthy or diseased, crop or not, or stressed or nonstressed) rather than its classes (i.e., disease type or crop name) to make a quick and efficient decision. Furthermore, ND can deal the best with unbalanced datasets compared to the other conventional classifiers (i.e., using the distribution of one class only to construct the model). Finally, the average classification rates of all subsets are passed to the ensemble stage (i.e., bagging in the proposed framework) to establish the final ensemble decision (using majority voting). The diversity is ensured and achieved by resampling input dataset and selecting and combining the best subsets of wavelengths from each subset.

The proposed classification framework can be described in four steps: spectral profile extraction, significant wavelength selection, ND classifier construction, and ensemble learning.

Step 1: Spectral profiles are extracted from the pixel values of a number of small areas from a leaf and then averaged over the entire wavelengths range. Manual or automatic process can be used in this stage to ensure that only useful leaf areas are selected. Furthermore, averaged reflectance is used to reduce the variation in pixel intensities across the leaf. It should be mentioned that the extracted profiles of the UoM dataset were scaled in order to minimize the variation in the profile of each condition using the following equation:

$$\rho_s = \frac{\rho - \rho_{\min}}{\rho_{\max} - \rho_{\min}} \quad (11)$$

where ρ , $\rho_{\min}$, $\rho_{\max}$, ρ_s represent measured, minimum, maximum, and scaled reflectance at each wavelength, respectively. Fig. 5 displays the scaled spectral profiles of averaged leaves with the standard errors. In addition, the ground truth reference map of the remotely sensed image was used to extract the spectral profile of the crops from the AVIRIS image.

Fig. 4. Proposed feature-ensemble-based novelty detection framework. Six feature selection algorithms (i.e., ReliefF, chi-square, Gini index, information gain, FCBF, and CFS) are used in each pool and only the best performing one is retained and used to extract the optimal subset of features. The classification accuracy of each pool is then calculated (using the optimal subset of each pool) and then passed to ensemble stage to contribute in the final classification accuracy (combined using majority voting).

Fig. 5. Scaled spectral profile. Average spectral profile of controlled, cold-stressed, and heat-stressed leaves with their standard error.

Step 2: Several feature selection algorithms {i.e., ReliefF, chi-square, Gini index, information gain, fast correlation based filter (FCBF), and CFS—detailed information about these algorithms can be found in our previous work [17]} are employed to preprocess the hyperspectral datasets, thus simplifying the datasets and selecting significant subset of wavelengths for each algorithm. This step is to prevent from overloading (i.e., dimension reduction), improve the prediction, and improve the robustness. However, these algorithms define the significance of the wavelengths based on different measuring criteria, and they differ

in the ability to reject irrelevant and redundant wavelengths. As a result, different algorithms generate different sets of wavelengths, and an overlap between the wavelengths may or may not be found. In addition, some redundant wavelengths may be included in the relevant subset of wavelengths due to the limitation of certain feature selection algorithms. It should be mentioned that there is no single feature selection algorithm that can deal with all situations, hence why various algorithms are used in our framework.

Step 3: ND-SVM is used for discrimination. The novelty score of the one-class SVM classifier [see (12)] is calibrated based on the measured stepwise-constant isotonic function to estimate the class probability, where a lower novelty score corresponds to a lower probability of being abnormal and vice versa. The probability is then used to define the threshold (i.e., taking the advantage of the probabilistic nature of the output) of the classifier as well as determine the best ND-SVM parameters (more details on calibration can be found in [39]):

$$z(\mathbf{x}) = b - \omega \cdot \phi(\mathbf{x}_i) \quad (12)$$

where b represents the bias, ω represents the weighting vector, and $\phi(\mathbf{x}_i)$ represents the transformed vector $\mathbf{x}_i$.

Note that the optimum solution of the dual problem [see (10)] can be determined if $\alpha_i > 0$ (i.e., $\frac{1}{\partial_t} > \alpha_i > 0$) and the weighting vector ω can be defined as

$$\omega = \sum_i \alpha_i \phi(\mathbf{x}_i) \quad (13)$$

as well as exploiting α_i , not at the lower or upper bounds, can be used to recover the bias:

$$b = (\omega \cdot \phi(\mathbf{x}_i)) = \sum_j \alpha_j K(\mathbf{x}_j, \mathbf{x}_i). \quad (14)$$

Novelty score is then used to classify test samples, whereas any score below the threshold is classified as normal and anything above as abnormal. The decision boundary is defined by (7), where its flexibility can be determined using the constraints of the support vectors presented in (10).

Step 4: In the ensemble stage, classification rates of resampled (k -fold) and preprocessed (i.e., selected wavelengths) of all subsets are passed to the ensemble to establish the final decision (or overall output) using majority voting. The overall classification performance of our framework has been compared to the existing empirical indices such as SR [83], modified chlorophyll absorption in reflectance index (MCARI) [84], transformed chlorophyll absorption in reflectance index (TCARI) [85], photochemical reflectance index (PRI) [86], disease water stress index 4 [57], and HI [53] as well as the classification performance of previous and published works [46], [47].

V. EXPERIMENTAL RESULTS

The experiments assessed the usefulness of the proposed framework for analyzing and classifying hyperspectral datasets. Three experiments were conducted: controlled vs. stressed, healthy vs. diseased, and one crop vs. other crops. The UoM dataset was used in the first experiment, the Bonn datasets (Bonn and Bonn-SPEC) were used in the second, and AVIRIS Indian Pines dataset was used in the third. Final results of the first two experiments were then compared with the results of the existing indices, while the last compared to the classification results reported in the literature. The CFS algorithm was used as a single feature selection algorithm in our experiments for the comparison purpose, due to the following reasons:

- 1) its ability to discard redundant or irrelevant features compared to other feature selection algorithms,
- 2) its applicability over a wide range of applications with different data characteristics,
- 3) easy to implement and to obtain optimal subset quickly, and
- 4) the number of wavelengths selected by CFS consistently lower than the most other methods with a satisfactory performance [25].

CFS uses the correlation between features and classes to evaluate the significance of the features. It uses Shannon's entropy $H(\mathbf{x}) = -\sum_{i=1}^n P(x_i) \log_2 P(x_i)$ and information gain $I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$ to measure average feature-class ($\bar{r}_{cf}$) and average feature-feature ($\bar{r}_{ff}$) correlations. The features are then evaluated heuristically on merit to determine the significance as follows:

$$\text{Merit}_S = \frac{n(\bar{r}_{cf})}{n + (n + n(n-1)\bar{r}_{ff})}. \quad (15)$$

The search strategy in this algorithm is based on the best first search and the process halts if no improvement is achieved after five consecutive iterations.

TABLE II
AVERAGE CLASSIFICATION RATES OF EMPIRICAL INDICES, ALL WAVELENGTHS, AND SELECTED WAVELENGTHS ON UoM DATASET

Technique	Average classification rate (%) (Standard deviation)
SR: R_{695}/R_{760}	68.51 (0.016)
MCARI	59.70 (0.028)
TCARI	55.84 (0.015)
PRI	51.85 (0.048)
DSSI-4	79.41 (0.010)
HI	78.74 (0.037)
ND-SVM (all wavelengths)	84.52 (0.020)
ND-SVM (CFS feature selection)	89.85 (0.018)
ND-SVM (feature ensemble)	91.58 (0.004)

A. Stress Detection

The UoM dataset, used in this experiment, consists of 216 controlled samples and 432 abnormal (i.e., cold and heat stressed) samples (i.e., 50% of each). A total of 60% of the normal (control) sample was used for training, and 40 percent plus all abnormal samples were used for testing and validation. Table II shows the averaged classification rate for 100 runs. The spectral profiles of each sample were interpolated (cubic) in order to extend the spectral resolution to 5 nm from 10 nm, so that the vegetation indices can be used. In addition, few spectral indices were used here, due to spectrum limitation (visible only).

What stands out in this table is that the improvement in average classification rate of the proposed method is more than 10% and 6% compared to the use of the best individual spectral index (DSSI-4 in this experiment) and all wavelengths, respectively. The literature has suggested a strong relationship between green pigment and stress detection [51]. The maximum sensitivity of green pigment that helps for early stress detection was found around 550–560 and 700–710 nm. Further analysis showed that different sets of wavelengths were selected by individual feature selection algorithms in the experiments; however, all of them selected three common wavelengths: 550, 680, and 710 nm; this closely conforms to the previous empirical study. The combination of these wavelengths with other wavelengths selected by individual algorithm leads to variations in the prediction performance and hence the improvements in the proposed framework. Further statistical tests revealed that the improvements are significant compared to using all wavelengths and single feature selection methods, at a significance level of 1% (p -value $< 10^{-5}$).

The ND experiment was repeated with the cold stress samples being considered as the normal class and all other samples as the abnormal. The experiment was also conducted with the heat stress being considered as the normal class and the rest as the abnormal. What is interesting is that the ND-SVM can also be used to detect different stresses and the results show significant improvements in the prediction performance compared to the empirical indices (87.14% with standard deviation of 0.026 with cold stress being the normal and 83.38% with standard deviation of 0.014 with heat stress being the normal).

In addition, the prediction performance of the ND-SVM was further compared with the two-class classifier. The SVM was used (one class, say control, against all others). In general, a two-class SVM performs slightly better than the ND-SVM setting with differences within 3.5%, due to the availability of other classes. However, the benefit of the ND-SVM setting is that only the normal samples are needed and used to construct the classification model. Moreover, the conventional SVM only outperforms our approach if it is used as a two-class classifier. In multiclass setting, i.e., all classes are classified separately, its average classification rate has dropped to 81.98% (standard deviation of 0.015), with 50% samples used as training and the remainders for testing with tenfold cross-validation.

B. Disease Detection

Spectral profiles of Bonn nonimaging and imaging systems were used in this experiment. Regarding nonimaging devices, the result of the novelty scores of the entire dataset (2516 samples with severity percentage starting from 0%) was compared with the result of the HI proposed in [53]. The procedure was then repeated with different disease severities (i.e., 1–5%). The ratio of the training samples to testing and validation in the normal data was 1.5:1 in all severity cases. With regard to the imaging system, 60% of the healthy samples (96 out of 160) was used to train our model, and the remaining (64 normal and 360 abnormal samples) were used to test and validate the model. Again, the classification rates, represented as average of 100 iterations, are presented in Tables III and IV.

Several studies have shown the potential of spectral vegetation and disease indices in detecting and identifying diseases. Closer inspection of Tables III and IV shows the advantage of the proposed framework compared to the existing indices and the use of the entire spectrum. Note that the samples of Bonn imaging dataset are sparsely distributed compared to the non-imaging samples. That is why the classification accuracy of all wavelengths outperformed the HI in the first four days (i.e., 7, 9, 11, and 13). Moreover, two thresholds were needed to differentiate healthy from diseased leaves if the value from HI was used for classification instead of a classifier. In addition, the discrimination performance of other disease indices was tested and showed poor separation ability. The result was not included since the interest here is to detect healthy species rather than diseased ones.

C. Crops Classification

The Indian Pines dataset was used in this experiment. The chief goal of this experiment was to check if the proposed framework can be applied to remotely sensed data and to assess its classification performance. The results were compared with the classification results published in the previous studies. Similar to the previous two experiments, 60% of the normal class was used for training and the remaining for testing and validation. Table V illustrates the average accuracy (AA) and the standard deviation.

This part of the experimental results has found that the proposed framework can play an important role in improving the

prediction performance of remote sensing classification, as confirmed in Table V. In comparison to the published results [46], [47], ensemble learning based bias-variance method and of combining SVMs and multiple classifier systems (i.e., called rotation-based SVM), the classification rate of the proposed framework displayed an improvement of 6.1% compared to all wavelengths, 3.6% compared to [46], and 2.8% compared to [47]. Note that no postprocessing was performed in our study, so the average rate was compared to the average rate of [46]. In addition, the average classification rate of our framework outperformed the conventional SVM reported in [46], [47].

VI. DISCUSSIONS

Several reports and articles have demonstrated the importance of the spectral indices in analyzing proximal and remotely sensed hyperspectral datasets; stress and disease detections [49], [51], [53], [56], [57]. Recently, many studies have highlighted the importance of feature extraction and selection techniques for classification [16], [18], [22], [60], to optimize a tradeoff between computational complexity and prediction performance [19]. A strong relationship between the number of features and ensemble performance has been reported in the literature [26]–[28], [66]. Moreover, several articles have highlighted the usefulness of combining feature selection and ensemble learning [17], [87]. In addition, recent attention has focused on the provision of robust and effective machine learning algorithms that can handle the one-class problems accurately. The current study provides evidence that a hyperspectral feature-ensemble framework with ND as a main classifier can improve the prediction performance of stress detection, disease detection, and crop classification. The focus is on normal samples where it is difficult to sample the entire abnormal cases.

The impact of feature selection on classifying HSI datasets was investigated, and the improvement was noticed. It is apparent from Tables II–V that very few features are needed to improve the classification rate, which agrees with our previous study [16]. The wavelengths selected by CFS and those used by empirical indices may share few common wavelengths, but the prediction performance of the former outperformed the latter. Poor discrimination of the indices might be due to the limited number of features used in prediction. It is now understood that systematic feature selection and combinations of the selected wavelengths can play an important role in enhancing discrimination ability. On the other hand, different feature selection algorithms generate different optimal sets, where part of the generated features in one algorithm might overlap with the others. In addition, different feature selection algorithms have inconsistent results in selecting features or prediction performance; this means that there is no single method to define the best algorithm for any specific dataset.

As discussed in the literature review, the performance of the combined learning algorithms is always preferable to the single learning algorithms [26]–[28], [66]. In addition, the importance of combining feature selection algorithms and ensemble learning was emphasized in several articles [17], [87]. The experimental results in the current study agree with these

TABLE III
AVERAGE CLASSIFICATION RATES OF HEALTHY INDEX, ALL WAVELENGTHS, AND SELECTED WAVELENGTHS ON BONN-SPEC DATASET WITH SEVERAL DISEASE SEVERITIES

Technique	Average classification rate (%) (Standard deviation)			
	0% Disease severity	1% Disease severity	3% Disease severity	5% Disease severity
HI	67.60 (0.011)	83.95 (0.002)	90.85 (0.004)	92.33 (0.007)
ND-SVM (all wavelengths)	62.96 (0.013)	79.21 (0.012)	88.02 (0.010)	91.17 (0.009)
ND-SVM (CFS feature selection)	78.75 (0.007)	85.59 (0.005)	91.55 (0.005)	93.70 (0.006)
ND-SVM (feature ensemble)	79.95 (0.013)	86.80 (0.013)	92.44 (0.009)	95.05 (0.008)

TABLE IV
AVERAGE CLASSIFICATION RATES OF HEALTHY INDEX, ALL WAVELENGTHS, AND SELECTED WAVELENGTHS ON BONN DATASET WITH DIFFERENT DISEASE INOCULATION DATES

Technique	Average classification rate (%) (Standard deviation)				
	Days 7–15	Days 9–15	Days 11–15	Days 13–15	Day 15
HI	66.35 (0.019)	70.48 (0.019)	79.68 (0.012)	85.92 (0.007)	92.66 (0.005)
ND-SVM (all wavelengths)	88.33 (0.013)	88.76 (0.015)	88.98 (0.015)	89.47 (0.020)	90.97 (0.022)
ND-SVM (CFS feature selection)	92.96 (0.014)	94.80 (0.012)	95.39 (0.015)	95.76 (0.019)	96.09 (0.025)
ND-SVM (feature ensemble)	93.90 (0.015)	95.32 (0.015)	97.12 (0.014)	98.23 (0.015)	98.52 (0.024)

TABLE V
AVERAGE CLASSIFICATION RATES OF SEVERAL CLASSIFICATION METHODS ON INDIAN PINES DATASET

Normal class	Average classification rate (%) (Standard deviation)				
	ND-SVM (all wavelengths)	ND-SVM (CFS)	ND-SVM (feature ensemble)	RoSVM [47]	RF [46]
1	99.89 (0.001)	99.56 (0.001)	99.85 (0.001)	100	
2	57.45 (0.011)	65.71 (0.011)	66.44 (0.014)	74.10	
3	58.03 (0.014)	68.40 (0.015)	82.78 (0.010)	76.32	
4	93.08 (0.009)	69.05 (0.041)	89.90 (0.012)	97.10	
5	92.05 (0.005)	82.26 (0.011)	94.62 (0.006)	93.87	
6	97.16 (0.001)	98.89 (0.001)	99.34 (0.001)	96.46	
7	98.79 (0.001)	99.95 (0.001)	99.95 (0.001)	100	
8	98.04 (0.003)	99.05 (0.001)	99.59 (0.001)	68.53	
9	98.81 (0.002)	99.90 (0.002)	99.94 (0.001)	100	
10	76.91 (0.016)	88.30 (0.016)	92.15 (0.007)	78.88	
11	64.44 (0.008)	72.42 (0.004)	77.71 (0.010)	65.69	
12	68.14 (0.021)	69.76 (0.018)	83.22 (0.012)	85.89	
13	99.01 (0.001)	99.74 (0.001)	99.84 (0.001)	99.42	
14	91.56 (0.002)	95.64 (0.002)	95.74 (0.002)	92.34	
15	76.96 (0.008)	81.29 (0.008)	86.23 (0.011)	92.78	
16	99.59 (0.001)	99.76 (0.001)	99.85 (0.001)	100	
AA (%)	85.62 (0.159)	86.86 (0.138)	91.70 (0.100)	88.84	88.07

AA stands for average accuracy. There were no class-specific accuracies reported in [46] for the RF method.

previous studies, where combining both techniques solves the inconsistency problem of feature selection algorithms and therefore leads to more robust discrimination. It should be noted that the feature-ensemble framework may not always beat the performance of the best individual feature selection algorithm. However, the combination is more robust, and overall it allows for better discrimination compared to any individual algorithm.

The literature on ND [39]–[41] has demonstrated the importance of such technique in solving unbalanced classification problem compared to conventional classifiers. Abnormal sam-

ples can be rare if not entirely absent or highly costly in many applications. Moreover, it is very difficult to cover all possible abnormal cases. Change detection in remotely sensed images is an example of the application of ND; the characteristics of the unchanged areas (usually available) are used during training in order to predict the changes.

The current study found that using feature ensemble with ND as a classifier can play an important role in setting a standard method for analyzing HSI datasets. Different theories exist in the literature regarding studying plant types and conditions, although they may not be applicable to all plant types and condi-

tions. Experimental results show an improvement of our framework in discrimination, 7% in stress detection, 3% in disease detection with different severity levels, 6% in disease detection with different inoculation dates, and approximately 3% in crop classification. A statistical test was performed to test the significance of our framework, and the result shows a significant improvement (at p -value $<10^{-5}$) in stress and disease detection cases. The findings show that the proposed framework is a valid approach to several applications including stress detection, disease detection, and crop classification from different sources. The relevance of this classification approach is manifold. For example, in the none-invasive real-time monitoring of abiotic plant stresses, such as droughting or salinization, it offers the potential to isolate the source, or sources, of stress in a more specific manner versus conventional gross stress indicators, such as stomatal conductance, as inferred from leaf canopy thermography, or photosynthesis reduction, from chlorophyll fluorescence. Similarly, for biotic stresses, such as the action of fungal pathogens, the close-proximity nature of the imaging data offers the potential to track the progress on stress symptoms across leaf and other plant tissue. The early use-cases for these classification approaches are most likely to be in rapid and wide area phenotype detection under laboratory and field conditions. The techniques also offer potential, beyond stress detection, within such applications as accurate mechanical control of crop conditions, notably within reduced or chemical-free management of persistent pests or resistant weeds, precise selective harvesting, notably in horticulture, or targeted nutrient feed directly to crop foliage.

VII. CONCLUSION

In this paper, a framework comprising of feature selection, novelty detection, and ensemble learning has been proposed for analysis and classification of HSI datasets of various sources with the primary goal of differentiating control and diseases plants, detecting stresses, and monitoring crop conditions. The experimental results showed marked improvements in the discrimination performance compared to the use of empirical vegetation indices and existing classification methods. The findings suggest the usefulness of the proposed framework for condition monitoring and abnormality detection when mostly only healthy samples are available, hence highly unbalanced data situation. In addition, the experiments across various HSI datasets have shown the validity and applicability of the proposed approach to a wide range of condition monitoring applications. Future work will explore the use of texture properties of HSI images in various spectral bands, together with the spectral information for studying more detailed changes in microstructural and pigment properties reflected in plant diseases and stresses for further enhanced analysis of their mechanisms.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. Johnson and Ms. H. Herrmann for the provision of Arabidopsis plants and Dr. K. Amano for helping the data collection with the UoM HSI system. They are also grateful to Dr. A.-K. Mahlein for the

provision of the Bonn datasets. The authors would also like to thank anonymous reviewers and the editor for their useful comments and suggestions, which helped improve the paper.

REFERENCES

- [1] A. F. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, no. 4704, pp. 1147–1153, 1985.
- [2] P. L. M. Geladi, H. F. Grahn, and J. E. Burger, "Multivariate images, hyperspectral imaging: Background and equipment," in *Techniques and Applications of Hyperspectral Image Analysis*. New York, NY, USA: Wiley, 2007, pp. 1–15.
- [3] G. ElMasry and D.-W. Sun, "Chapter 1: Principles of hyperspectral imaging technology," in *Hyperspectral Imaging for Food Quality Analysis and Control*, D.-W. Sun, Ed. San Diego, CA, USA: Academic, 2010, pp. 3–43.
- [4] J. Campbell and R. Wynne, *Introduction to Remote Sensing*, 5th ed. New York, NY, USA: Guilford Publications, 2011.
- [5] A. Cracknell and L. Hayes, *Introduction to Remote Sensing*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2007.
- [6] Y. Zhang, D. C. Slaughter, and E. S. Staab, "Robust hyperspectral vision-based classification for multi-season weed mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 69, pp. 65–73, 2012.
- [7] H. Liu, S.-H. Lee, and J. S. Chahl, "Development of a proximal machine vision system for off-season weed mapping in broadacre no-tillage fallows," *J. Comput. Sci.*, vol. 9, no. 12, pp. 1803–1821, 2013.
- [8] P. Geladi, E. Bengtsson, K. Esbensen, and H. Grahn, "Image analysis in chemistry I. Properties of images, greylevel operations, the multivariate image," *Trends Anal. Chem.*, vol. 11, no. 1, pp. 41–53, 1992.
- [9] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, 2014, Art. no. 010 901.
- [10] C. Duchesne, J. Liu, and J. MacGregor, "Multivariate image analysis in the process industries: A review," *Chemometrics Intell. Lab. Syst.*, vol. 117, pp. 116–128, 2012.
- [11] E. Barrett and L. Curtis, *Introduction to Environmental Remote Sensing*. Dordrecht, The Netherlands: Springer, 1992.
- [12] J. Qin, "Chapter 5: Hyperspectral imaging instruments," in *Hyperspectral Imaging for Food Quality Analysis and Control*, D.-W. Sun, Ed. San Diego, CA, USA: Academic, 2010, pp. 129–172.
- [13] I. Herrmann, U. Shapira, S. Kinast, A. Karnieli, and D. J. Bonfil, "Ground-level hyperspectral imagery for detecting weeds in wheat fields," *Precis. Agriculture*, vol. 14, no. 6, pp. 637–659, 2013.
- [14] A. AlSuwaidi, "Dubaisat-2 mission overview," *SPIE Remote Sens.*, vol. 8533, 2012, Art. no. 85330W.
- [15] A. AlRais, A. AlSuwaidi, and H. Ghedira, "Dubaisat-1: Mission overview, development status and future applications," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 5, Jul. 2009, pp. V-196–V-199.
- [16] A. AlSuwaidi, C. Veys, M. Hussey, B. Grieve, and H. Yin, "Hyperspectral selection based algorithm for plant classification," in *Proc. IEEE Int. Conf. Imag. Syst. Techn.*, Oct. 2016, pp. 395–400.
- [17] A. AlSuwaidi, C. Veys, M. Hussey, B. Grieve, and H. Yin, "Hyperspectral feature selection ensemble for plant classification," in *Proc. Hyperspectr. Imag. Appl.*, Session 4-12.00, Oct. 2016.
- [18] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
- [19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2000.
- [20] L. Jiang, B. Zhu, and Y. Tao, "Chapter 3: Hyperspectral image classification methods," in *Hyperspectral Imaging for Food Quality Analysis and Control*, D.-W. Sun, Ed. San Diego, CA, USA: Academic, 2010, pp. 79–98.
- [21] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA, USA: Kluwer, 1998.
- [22] L. C. Molina, L. Belanche, and A. Nebot, "Feature selection algorithms: A survey and experimental evaluation," in *Proc. IEEE Int. Conf. Data Mining*, 2002, pp. 306–313.
- [23] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. 10th Nat. Conf. Artif. Intell.*, 1992, pp. 129–134.
- [24] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of Relief and RRelief," *Mach. Learn.*, vol. 53, no. 1, pp. 23–69, 2003.
- [25] M. A. Hall and L. A. Smith, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proc. 12th Int. Florida Artif. Intell. Res. Soc. Conf.*, 1999, pp. 235–239.

- [26] I. H. Witten, E. Frank, and M. A. Hall, "Chapter 8: Ensemble learning," in *Data Mining: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems)*, I. H. Witten, E. Frank, and M. A. Hall, Eds., 3rd ed. Boston, MA, USA: Morgan Kaufmann, 2011, pp. 351–373.
- [27] T. G. Dietterich, "Machine-learning research: Four current directions," *Artif. Intell. Mag.*, vol. 18, pp. 97–136, 1997.
- [28] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Third Quart. 2006.
- [29] L. O. Jimenez, A. Morales-Morell, and A. Creus, "Classification of hyper-dimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1360–1366, May 1999.
- [30] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," *IETE Tech. Rev.*, vol. 27, no. 4, pp. 293–307, 2010.
- [31] Y. Zhong, Q. Cao, J. Zhao, A. Ma, B. Zhao, and L. Zhang, "Optimal decision fusion for urban land-use/land-cover classification based on adaptive differential evolution using hyperspectral and LiDAR data," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 868.
- [32] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.
- [33] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
- [34] P. Ghamisi, A. R. Ali, M. S. Couceiro, and J. A. Benediktsson, "A novel evolutionary swarm fuzzy clustering approach for hyperspectral imagery," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2447–2456, Jun. 2015.
- [35] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 5, pp. 1087–1095, Sep. 1994.
- [36] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6937–6956, Nov. 2014.
- [37] L. Hamel, "Novelty detection," in *Knowledge Discovery With Support Vector Machines*. New York, NY, USA: Wiley, 2009, pp. 209–217.
- [38] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, 2014.
- [39] L. Clifton, D. A. Clifton, Y. Zhang, P. Watkinson, L. Tarassenko, and H. Yin, "Probabilistic novelty detection with support vector machines," *IEEE Trans. Reliab.*, vol. 63, no. 2, pp. 455–467, Jun. 2014.
- [40] F. de Morsier, D. Tuia, M. Borgeaud, V. Gass, and J. P. Thiran, "Semi-supervised novelty detection using SVM entire solution path," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 1939–1950, Apr. 2013.
- [41] D. A. Clifton, L. A. Clifton, P. R. Bannister, and L. Tarassenko, "Automated novelty detection in industrial systems," in *Advances of Computational Intelligence in Industrial Systems*, Y. Liu, A. Sun, H. T. Loh, W. F. Lu, and E.-P. Lim, Eds. Berlin, Germany: Springer, 2008, pp. 269–296.
- [42] W. Fisher, "Progress in remote sensing (1972–1976)," *Photogrammetria*, vol. 32, pp. 33–72, 1976.
- [43] F. A. Kruse, J. W. Boardman, and J. F. Huntington, "Comparison of airborne hyperspectral data and EO-1 Hyperion for mineral mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1388–1400, Jun. 2003.
- [44] H. Breit, T. Fritz, U. Balss, M. Lachaise, A. Niedermeier, and M. Vonavka, "TerraSAR-X SAR processing and products," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 727–740, Feb. 2010.
- [45] A. AlSuwaidi, A. AlRais, and H. Ghedira, "Developing a satellite-based tool to monitor dust and sand storms in the UAE," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2010, pp. 1434–1437.
- [46] A. Merentitis, C. Debes, and R. Heremans, "Ensemble learning in hyperspectral image classification: Toward selecting a favorable bias-variance tradeoff," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1089–1102, Apr. 2014.
- [47] J. Xia, J. Chanussot, P. Du, and X. He, "Rotation-based support vector machine ensemble in classification of hyperspectral data with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1519–1531, Mar. 2016.
- [48] K. H. Holland, D. W. Lamb, and J. S. Schepers, "Radiometry of proximal active optical sensors (AOS) for agricultural sensing," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 6, pp. 1793–1802, Dec. 2012.
- [49] D. A. Sims and J. A. Gamon, "Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages," *Remote Sens. Environ.*, vol. 81, no. 23, pp. 337–354, 2002.
- [50] G. Rondeaux, M. Steven, and F. Baret, "Optimization of soil-adjusted vegetation indices," *Remote Sens. Environ.*, vol. 55, no. 2, pp. 95–107, 1996.
- [51] A. Gitelson and M. N. Merzlyak, "Spectral reflectance changes associated with autumn senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. leaves. spectral features and relation to chlorophyll estimation," *J. Plant Physiol.*, vol. 143, no. 3, pp. 286–292, 1994.
- [52] T. P. Dawson and P. J. Curran, "A new technique for interpolating the reflectance red edge position," *Int. J. Remote Sens.*, vol. 19, no. 11, pp. 2133–2139, 1998.
- [53] A.-K. Mahlein, T. Rumpf, P. Welke, H.-W. Dehne, L. Plmer, U. Steiner, and E.-C. Oerke, "Development of spectral indices for detecting and identifying plant diseases," *Remote Sens. Environ.*, vol. 128, pp. 21–30, 2013.
- [54] J. Penuelas, J. Pinol, R. Ogaya, and I. Filella, "Estimation of plant water concentration by the reflectance water index WI (R900/R970)," *Int. J. Remote Sens.*, vol. 18, no. 13, pp. 2869–2875, 1997.
- [55] B.-C. Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sens. Environ.*, vol. 58, no. 3, pp. 257–266, 1996.
- [56] T. Rumpf, A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, and L. Plmer, "Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance," *Comput. Electron. Agriculture*, vol. 74, no. 1, pp. 91–99, 2010.
- [57] A. Apan, A. Held, S. Phinn, and J. Markley, "Detecting sugarcane orange rust disease using EO-1 Hyperion hyperspectral imagery," *Int. J. Remote Sens.*, vol. 25, no. 2, pp. 489–498, 2004.
- [58] S. L. Ustin, D. A. Roberts, M. Gardner, and P. Dennison, "Evaluation of the potential of Hyperion data to estimate wildfire hazard in the Santa Ynez front range, Santa Barbara, California," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 2, 2002, pp. 796–798.
- [59] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
- [60] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.
- [61] A. M. De Silva and P. H. W. Leong, "Feature selection," in *Grammar-Based Feature Generation for Time-Series Prediction*. Singapore: Springer, 2015, pp. 13–24.
- [62] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujn, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, 2012.
- [63] J. Reiche, C. M. Souza, D. H. Hoekman, J. Verbesselt, H. Persaud, and M. Herold, "Feature level fusion of multi-temporal ALOS PALSAR and Landsat data for mapping and monitoring of tropical deforestation and forest degradation," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 5, pp. 2159–2173, Oct. 2013.
- [64] W. Liao, A. Piurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 552–556, Mar. 2015.
- [65] X. Bai, C. Liu, P. Ren, J. Zhou, H. Zhao, and Y. Su, "Object classification via feature fusion based marginalized kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 8–12, Jan. 2015.
- [66] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *J. Inf. Fusion*, vol. 6, pp. 5–20, 2005.
- [67] S. Prasad and L. M. Bruce, "Decision fusion with confidence-based weight assignment for hyperspectral target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1448–1456, May 2008.
- [68] A. C. Turlapaty, H. K. Gori, and B. Gokaraju, "A feature subset based decision fusion approach for scene classification using color, spectral, and texture statistics," in *Proc. IEEE 7th Int. Adv. Comput. Conf.*, Jan. 2017, pp. 624–628.
- [69] D. M. Tax and R. P. Duin, "Support vector data description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.
- [70] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.
- [71] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer, 2006.
- [72] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using K-nearest neighbour graph," in *Proc. 7th Int. Conf. Pattern Recog.*, vol. 3, Aug. 2004, pp. 430–433.

- [73] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.*, 1999, pp. 582–588.
- [74] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.
- [75] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Hoboken, NJ, USA: Wiley-Interscience, 2006.
- [76] S. Kulkarni and G. Harman, *An Elementary Introduction to Statistical Learning Theory*, 1st ed. New York, NY, USA: Wiley, 2011.
- [77] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
- [78] A.-K. Mahlein, U. Steiner, H.-W. Dehne, and E.-C. Oerke, "Spectral signatures of sugar beet leaves for the detection and differentiation of diseases," *Precis. Agriculture*, vol. 11, no. 4, pp. 413–431, 2010.
- [79] D. H. Foster, K. Amano, and S. M. C. Nascimento, "Color constancy in natural scenes explained by global image statistics," *Vis. Neurosci.*, vol. 23, nos. 3–4, pp. 341–349, 2006.
- [80] D. H. Foster, K. Amano, S. M. C. Nascimento, and M. J. Foster, "Frequency of metamerism in natural scenes," *J. Opt. Soc. Amer. A*, vol. 23, no. 10, pp. 2359–2372, Oct. 2006.
- [81] A.-K. Mahlein, U. Steiner, C. Hillnhütter, H.-W. Dehne, and E.-C. Oerke, "Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases," *Plant Methods*, vol. 8, no. 1, 2012, Art. no. 3.
- [82] A.-K. Mahlein, S. Hammersley, E.-C. Oerke, H.-W. Dehne, H. Goldbach, and B. Grieve, "Supplemental blue LED lighting array to improve the signal quality in hyperspectral imaging of plants," *Sensors*, vol. 15, no. 6, pp. 12834–12840, 2015.
- [83] G. A. Carter, "Ratios of leaf reflectances in narrow wavebands as indicators of plant stress," *Int. J. Remote Sens.*, vol. 15, no. 3, pp. 697–703, 1994.
- [84] C. Daughtry, C. Walthall, M. Kim, E. de Colstoun, and J. McMurtry III, "Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance," *Remote Sens. Environ.*, vol. 74, no. 2, pp. 229–239, 2000.
- [85] D. Haboudane, J. R. Miller, N. Tremblay, P. J. Zarco-Tejada, and L. Dextraze, "Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture," *Remote Sens. Environ.*, vol. 81, no. 23, pp. 416–426, 2002.
- [86] J. Gamon, J. Peuelas, and C. Field, "A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency," *Remote Sens. Environ.*, vol. 41, no. 1, pp. 35–44, 1992.
- [87] G.-Z. Li and J. Y. Yang, "Feature selection for ensemble learning and its application," in *Machine Learning in Bioinformatics*. New York, NY, USA: Wiley, 2008, pp. 135–155.

Ali AlSuwaidi (S'06) received the B.Eng. degree in communication engineering in 2007 from the Khalifa University (formerly the Etisalat University College), Abu Dhabi, United Arab Emirates, and the M.Sc. degree in digital image and signal processing in 2011 from The University of Manchester, Manchester, U.K., where he has been working toward the Ph.D. degree in electrical and electronic engineering since 2014.

He worked with the Emirates Institution for Advanced Science and Technology (currently Mohammed Bin Rashid Space Centre), from 2008 to 2014, as the Head of the Applications Development and Analysis Centre. He participated in DubaiSat-1 and DubaiSat-2 Launches and was also involved in a number of research activities where he developed a number of tools for environmental applications based on satellite imagery and spatial data. His main research interests include pattern recognition, machine learning, artificial intelligence, image and signal processing, and remote and proximal sensing.

Bruce Grieve received the BSc (Hons) degree in chemical engineering from the University of York in 1988 and the PhD degree in process engineering from the University of Manchester in 2002. He gained 18 years of industrial experience in the fields of online analysis and measurement R&D, including deployment of sensors and informatics systems within new integrated products for sustainable agriculture and food, before joining The University of Manchester, Manchester, U.K., as the Director of the e-Agri Sensors Centre. Since 2007, he has attracted more than 5.3M of direct industrial, governmental, and NGO funding, which has been leveraged against a portfolio of 20.4M in collaborative multidisciplinary research projects. Previously, he was the Industrial Manager on a number of U.K. Research Council and DTI supported projects.

Prof. Grieve is a Fellow of the Institute of Engineering and Technology, a Fellow of the Institute of Agricultural Engineers, and a Fellow of the Higher Education Academy, and holds the N8 Chair in Agri-Sensors and Electronics. He was the recipient of the Royal Academy of Engineering Senior Fellowship to progress his e-Agri research at The University of Manchester and has held a number of funding board roles with U.K. Research Councils and the Innovate-UK (TSB), including being a nominated member of the BBSRC Agriculture and Food Security Strategy and Policy Panel and the STFC 21st Century Challenges Strategy Panel.

Hujun Yin (SM'03) received the B.Eng. degree in electronic engineering and the M.Sc. degree in signal processing from Southeast University, Nanjing, China, in 1983 and 1986, respectively, and the Ph.D. degree in neural networks from the University of York, York, U.K. in 1996.

He has been with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester, U.K., since 1998. He has studied and extended the self-organizing neural networks and related unsupervised learning topics, such as manifolds, dimensionality reduction, and data visualization, and proposed a number of methods for associated data analytics and modeling. He has authored or co-authored more than 200 peer-reviewed articles in a range of topics, from neural networks, density modeling, image processing, face recognition, text mining and knowledge management, gene expression analysis and peptide sequencing, novelty detection, to financial time series modeling, and recently decoding neuronal responses. His main research interests include neural networks, self-organizing systems, deep learning, image and hyperspectral image processing, face recognition, and bio-/neuro-informatics.

Dr. Yin is a Member of the U.K. EPSRC College. He is an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS and a Member of the Editorial Board of the *International Journal of Neural Systems*. He had also served as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS between 2006 and 2010. He has been the Organizing Chair, Programme Committee Chair, and General Chair for several international conferences, such as International Workshop on Self-Organizing Maps, International Conference on Intelligent Data Engineering and Automated Learning, and International Symposium on Neural Networks.