Chemometrics and Intelligent Laboratory Systems 217 (2021) 104404

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Estimation of nitrogen content in cucumber plant (*Cucumis sativus* L.) leaves using hyperspectral imaging data with neural network and partial least squares regressions

Sajad Sabzi [a], Razieh Pourdarbani [b,**], Mohammad H. Rohban [a], Ginés García-Mateos [c], Juan I. Arribas [d,e,*]

[a] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[b] Department of Biosystems Engineering, University of Mohaghegh Ardabili, Ardabil, Iran
[c] Department of Computer Science and Systems, University of Murcia, 30100, Murcia, Spain
[d] Department of Electrical Engineering, University of Valladolid, 47011, Valladolid, Spain
[e] Castilla-Leon Neuroscience Institute, University of Salamanca, 37007, Salamanca, Spain

ARTICLE INFO

Keywords:
Cucumber
Hyperspectral imaging
Image processing
Leaf
Machine learning
Nitrogen
Optimization
Plant
Prediction
Regression

ABSTRACT

In recent years, farmers have often mistakenly resorted to overuse of chemical fertilizers to increase crop yield. However, excessive consumption of fertilizers might lead to severe food poisoning. If nutritional deficiencies are detected early, it can help farmers to design better fertigation practices before the problem becomes unsolvable. The aim of this study is to predict the amount of nitrogen (N) content (mg l^{-1}) in cucumber (*Cucumis sativus* L., var. Super Arshiya-F1) plant leaves using hyperspectral imaging (HSI) techniques and three different regression methods: a hybrid artificial neural networks-particle swarm optimization (ANN-PSO); partial least squares regression (PLSR); and unidimensional deep learning convolutional neural networks (CNN). Cucumber plant seeds were planted in 20 different pots. After growing the plants, pots were categorized and three levels of nitrogen overdose were applied to each category: 30%, 60% and 90% excesses, called $N_{30\%}$, $N_{60\%}$, $N_{90\%}$, respectively. HSI images of plant leaves were captured before and after the application of nitrogen excess. A prediction regression model was developed for each individual category. Results showed that mean regression coefficients (R) for ANN-PSO were inside 0.937–0.965, PLSR 0.975–0.997, and CNN 0.965–0.985 ranges, test set. We conclude that regression models have a remarkable ability to accurately predict the amount of nitrogen content in cucumber plants from hyperspectral leaf images in a non-destructive way, being PLSR slightly ahead of CNN and ANN-PSO methods.

1. Introduction

In recent years, agricultural producers have increased the application of chemical fertilizers per unit area, instead of using modern agricultural knowledge to increase production. The fallacy of increased yield due to increased use of water and chemical fertilizers has led to excessive application of fertilizer resources. However, continuation over time of these actions may pose serious risks of contamination and threat to human health, not to mention financial losses and intensification of nutrient imbalances in soil [1,2]. For example, nitrogen (N) fertilizer, which is consumed in large quantities due to its low cost, is converted to nitrate being a known carcinogen with the potential of causing endocrine and immune systems disorders which might lead to gastrointestinal cancer [3].

The first symptom of nitrogen poisoning in cucumber plants includes the appearance of a yellow spot on the surface of plant leaves, which subsequently expands so that it may cover the entire leaf surface, except where its veins remain green. In most cases, nitrogen poisoning can be eliminated by heavy irrigation and proper environmental control. Thus, early detection of plant poisoning may help to reduce the severity of this problem as soon as possible [4].

Plant leaves are an important source of information for plant recognition, identification and health monitoring. The development of intelligent computer systems based on leaf analysis can help in non-

* Corresponding author. Department of Electrical Engineering, University of Valladolid, 47011, Valladolid, Spain.
** Corresponding author.
E-mail addresses: r_pourdarbani@uma.ac.ir (R. Pourdarbani), jarribas@tel.uva.es (J.I. Arribas).

https://doi.org/10.1016/j.chemolab.2021.104404
Received 31 March 2021; Received in revised form 11 August 2021; Accepted 13 August 2021
Available online 17 August 2021

Table 1
Statistical analysis of ANOVA to evaluate the significance difference between spectral data of categories.

	Sum of Squares	Degrees of freedom	Mean Square	F-test	Significance
Between Groups	2.015E11	3	6.715E10	723.600	.000*
Within Groups	3.666E10	395	9.280E7		
Total	2.381E11	398			

Table 2
Duncan test for evaluating the significance difference between spectral data categories.

Pot	n	Control pot	Excess N (30%)	Excess N (60%)	Excess N (90%)
Excess N (90%)	100	4.9128E3			
Excess N (30%)	100		5.1189E4		
Excess N (60%)	100			5.4648E4	
Control pot	100				6.2398E4
Significance		1.000	1.000	1.000	1.000

Table 3
Structure of the hidden layers of the ANN-SA to select the most effective spectral wavelengths. *radbas*: radial basis function. *trainrp*: resilient backpropagation algorithm. *learnh*: Hebb learning function.

Number of hidden layers	Number of neurons per layer	Transfer function	Back-propagation training function	Back-propagation weight learning/ bias
2	12 & 17	*radbas*	*trainrp*	*learnh*

Table 4
Structure of the convolutional neural network (CNN) used to estimate nitrogen content in cucumber plant leaves. Input is a (327 × 1) vector of light absorbance values corresponding to spectral sample wavelengths. The total number of parameters of the network is 551,617.

Layer type	Convolution size	Output shape (width × features)	Number of params.
Convolution + ReLu	7 × 1	321 × 32	256
Max pooling		160 × 32	–
Convolution + ReLu	5 × 32	156 × 64	10304
Max pooling		78 × 64	–
Convolution + ReLu	5 × 64	74 × 128	41088
Max pooling		37 × 128	–
Convolution + ReLu	3 × 128	35 × 256	98560
Max pooling		17 × 256	–
Convolution + ReLu	3 × 256	15 × 512	393728
Flatten		7680	–
Dense		1	7681

destructive and real-time assessment of plants [5]. So far, several methods have been developed for non-destructive prediction of agricultural products, using plant leaves. With the advancement of machine vision and machine learning, significant progress has been made recently in recognizing plant characteristics [6–8]. The performance of these methods depends on two factors: a) the accuracy of these methods depends on the extraction and selection of measurable (often visual) features; and b) noise in the images that is largely unavoidable under real farm conditions. Several types of technologies have been used so far to identify products, including electrical resistance spectroscopy [9], reflectance spectroscopy [10], Fourier infrared spectroscopy [11] and fluorescence spectroscopy [12].

Among the most promising techniques for non-destructive monitoring of plant leaves properties, we can find many recent advances in hyperspectral imaging (HSI). HSI integrates both spectroscopic and imaging techniques into a system capable of computing a spatial map of the spectral content variations [13–15]. For example, Serranti et al. [16] monitored fertilizer production process using HSI. They studied different mixtures of urban organic waste to find correlations between parameters such as pH, electrical conductivity, soluble nitrogen and soluble carbon. The results revealed a correlation coefficient between 0.85 and 0.96 for all parameters, indicating the ability of HSI for non-destructive monitoring of waste-derived fertilizer production. More recently, Zhou et al. [17] predicted the germination of sugar beet seeds using HSI with an accuracy of 89% over the test set.

Several other works can be found in the literature concerning the estimation of the chemical properties of the plants using spectroscopy, and more specifically HSI. As an example, Eshkabilov et al. [18] studied on the detection of the level of nutrients in lettuce. Partial least squares regression (PLSR) and principal component analysis (PCA) were used. They captured hyperspectral images of lettuce leaves using a hyperspectral camera at wavelength range of 400–1000 nm. The most effective wavelengths were found at 506–601 nm and 634–701 nm ranges. Nutrient content including soluble solid content, pH, nitrate, calcium, potassium and total chlorophyll, were predicted by both PLSR and PCA optimization algorithms. Results showed a determination coefficient (R^2) in the range of 0.784–0.987 in predicting nutrients content. Closely related to our problem, Chen et al. [19] predicted nitrogen content in apple-trees using HSI. Original wavelengths were preprocessed by different smoothing filters, including multiplicative scatter correction (MSC), Savitzky-Golay (SG) smoothing, normalization by the mean (NME), standard normal transformation (SNV), and first-order derivative (FD) or second-order derivative (SD). Results showed that SNV-FD had the best performance. Then, features were extracted by the successive projection algorithm (SPA), random frog (Rfrog), and partial least squares (PLS). According to the results, Rfrog-Extreme Learning Machine (ELM) had the best accuracy, with an R^2 of 0.843, RMSE 2.461 g kg^{-1}, and relative percent deviation RPD of 2.508.

Other types of chemical elements have also been analyzed using HSI technologies. Wang et al. [20] estimated Sodium (K) and Potassium (P) content in tea leaves using HSI and chemometrics. The key spectrum wavelengths were selected by SPA algorithm, and regression coefficients (RC) of PLSR model. Successive projections algorithm-multiple linear regression (SPA-MLR) achieved a good performance with correlation coefficients of 0.9423 for P, and 0.9168 for K. Sun et al. [21] accomplished a study on the estimation of water content in corn leaves by hyperspectral images. Again related to nitrogen content, Liu et al. [22] used aerial hyperspectral data to model nitrogen in potato at three growth stages and two growing seasons. The results revealed that short-wave infrared (SWIR) wavelengths are important for modelling both crop variables and either cultivar or season-specific models, but showing a systematic prediction bias. Amoah et al. [23] treated freezing injury treatment in tea leaves with nitrogen and assessed the effect of nitrogen using HSI, analyzed with PLSR, Principal Component Regression, and Linear Models. They concluded that the wavelength at which absorption bands occur can influence the model performance significantly.

As can be inferred from previous introduction, hyperspectral images can be used for reliable diagnosis of the nutritional status of agricultural products, obtaining accurate and powerful tools in farm managing. In this context, the present study aims to predict excess nitrogen content in cucumber plants using HSI data from plant leaves, in a non-destructive way. First, the most effective wavelengths for the case of study are extracted using a machine learning method. Then, the information in the selected wavelengths is analyzed with three regression methods (a hybrid artificial neural networks and the particle swarm optimization (ANN-PSO), partial least squares regression (PLSR), and a deep learning

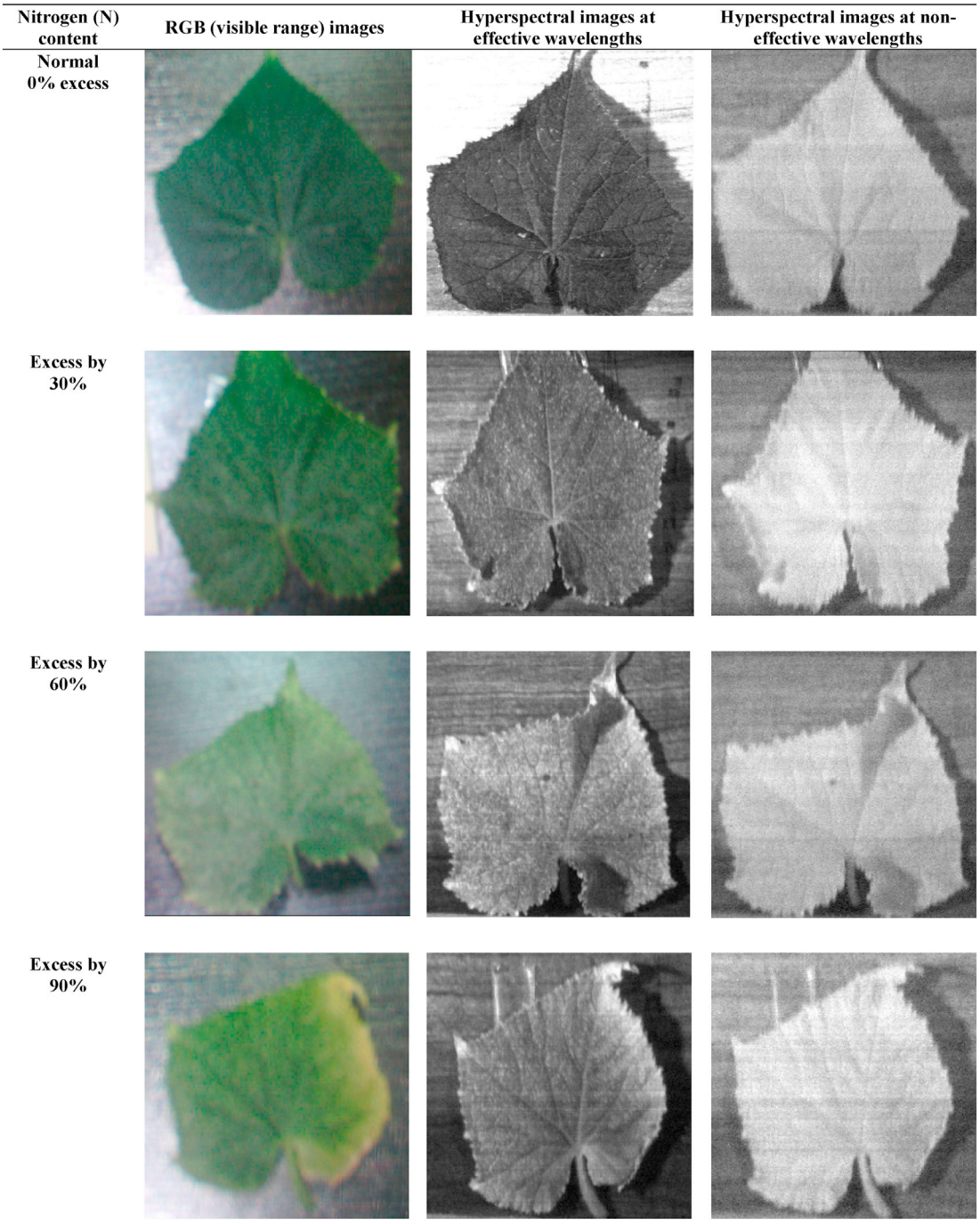

Fig. 1. Examples of cucumber plant leaves images, including visible range RGB **(left)** and both effective (optimal) **(center)** and not-effective **(right)** hyperspectral wavelengths. Rows indicate nitrogen treatment class: normal, 30%, 60%, and 90% nitrogen excess.

method based on convolutional neural networks (CNN)) in order to detect potential nitrogen excessive levels as early as possible.

2. Materials and methods

A global outline of the research procedure applied in the present study is depicted in Fig. S1. The process begins with data acquisition, which includes plant cultivation, application of different fertilizer contents, and HSI acquisition of cucumber plant leaves. Data is analyzed, producing either a joint prediction model or individual models for each treatment depending on their statistical significance differences. Then, the most effective wavelengths are extracted using a machine learning approach. Finally, three regression methods are applied and compared for the estimation of nitrogen (N) content in plant leaves: ANN-PSO, PLSR and CNN.

2.1. Data collection

Cucumber seeds (Super Arshiya-F1 variety) were planted in 20 pots, as shown in Fig. S2a, to prepare samples with standard nitrogen and nitrogen excess contents. The fertilizer used was a compound of urea and ammonium, and it was applied with irrigation water at a standard concentration of 2 g/l. Whenever leaves were grown, approximately 3 months after planting, pots were categorized into four groups: pots with standard nitrogen content (0% N excess), and pots with applied nitrogen excess by 30%, 60% and 90%, respectively. First, 50 leaves were

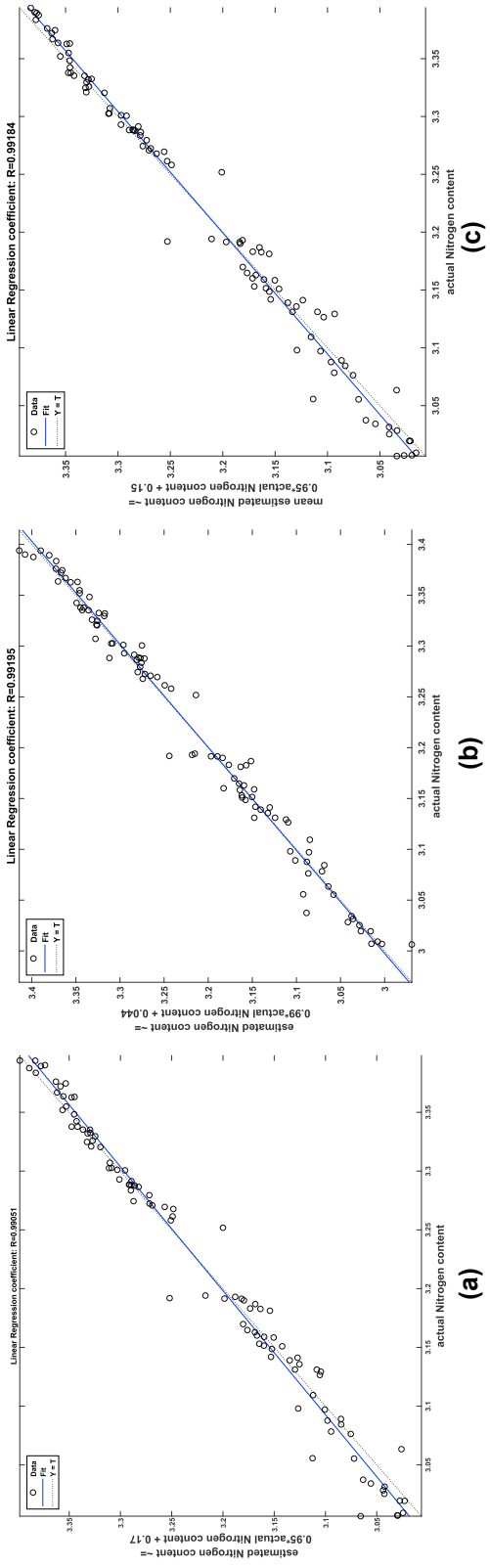

Fig. 2. Scatter plot of regression analysis for mean estimated and measured nitrogen content (mg l^{-1}), in the treatment of nitrogen excess by 30%, after 100 iterations (test set): (a) ANN-PSO, (b) PLSR, and (c) CNN.

collected from the first group (control pot) and were imaged by a hyperspectral camera. A total of 24 h after the application of nitrogen excess, 50 leaves from each of the three plant categories were picked and imaged, resulting 100 plant leaves for each N content plant category (totaling 300 plant leaves images). Hyperspectral images were also captured 48 h and 72 h after the over-application of N. Since the ultimate goal of prediction was to early recognize nitrogen-rich leaves, the data related to the first 24 h after nitrogen application were analyzed first. In the event that results were not acceptable (accurate enough), the remainder plant leaves data would be entered into the system (48 h and 72 h).

2.2. Extraction of spectral information from plant leaves

In order to extract spectral information from each leaf, a hyperspectral camera (Physics Noor Co., Fanavaran, Kashan, Iran) in the visible and near-infrared (Vis/NIR) range (400–1000 nm) was used. The HSI camera (shown in Fig. S2b) was placed inside an illumination chamber (Fig. S2c) equipped with two tungsten halogen lamps SLI-CAL (StellarNet, Tampa, Florida, USA) to prevent disturbing ambient light. The samples were located at a horizontal distance of 1 m from the camera and were illuminated by both lamps. Also, a conventional laptop (Intel Corei5, 2430 M at 2.40 GHz, 4 GB of RAM, Windows 10) was used to store and process the data.

The HSI camera captured a total of 327 images from each plant leaf, each of them corresponding to an specific wavelength inside 400–1000 nm range. Thus, these images contain the reflectance information of the samples at each point. Rossel [24] observed that using absorption spectra is more adequate than reflection data, since the former presents a linear relationship with the molecular concentration of the samples. Consequently, images were converted to absorption information using the following equation:

$$B\left(x,y\right) = \log\left({}^{1}/_{A(x,y)}\right) \tag{1}$$

where A is input image (reflectance) data, B the resulting output image (absorption) data, and (x,y) are all pixels inside a two-dimensional image.

2.3. Extraction of nitrogen (N) content using laboratory analysis

Besides hyperspectral data, the actual nitrogen (N) content of plant leaves was also measured using a standard chemical procedure [25]. First, leaves were dried and pulverized. To determine total nitrogen, a Gerhardt Vap20 digester (Gerhardt GmbH & Co., Königswinter, Germany) was used. After the digestion step, refrigerant was used to distill the sample and finally titrated. After the titration step, the total nitrogen content was obtained by calculating the consumption of acid (c.f. Fig. S3). For this purpose, equation (2) is used:

$$N_{\,total} = \frac{(vs - vb)}{md} \times N_{H_2SO_4} \times 0.014\ meq\ N \times 100 \tag{2}$$

where vs is the consumed volume of the sample, vb is the consumed volume of the control treatment, $N_{H_2SO_4}$ is the normality of sulphuric acid at 0.014 meq, and md is the dry weight of the sample. Measured nitrogen values are used as the ground-truth for the three proposed regression methods.

2.4. Statistical analysis of the spectral data

As early presented in the schematic Fig. S1, before developing a regression model for the prediction of nitrogen (N) content in cucumber plant leaf images, the entire data should be statistically examined to observe the differences between the treatments. If there is a statistically

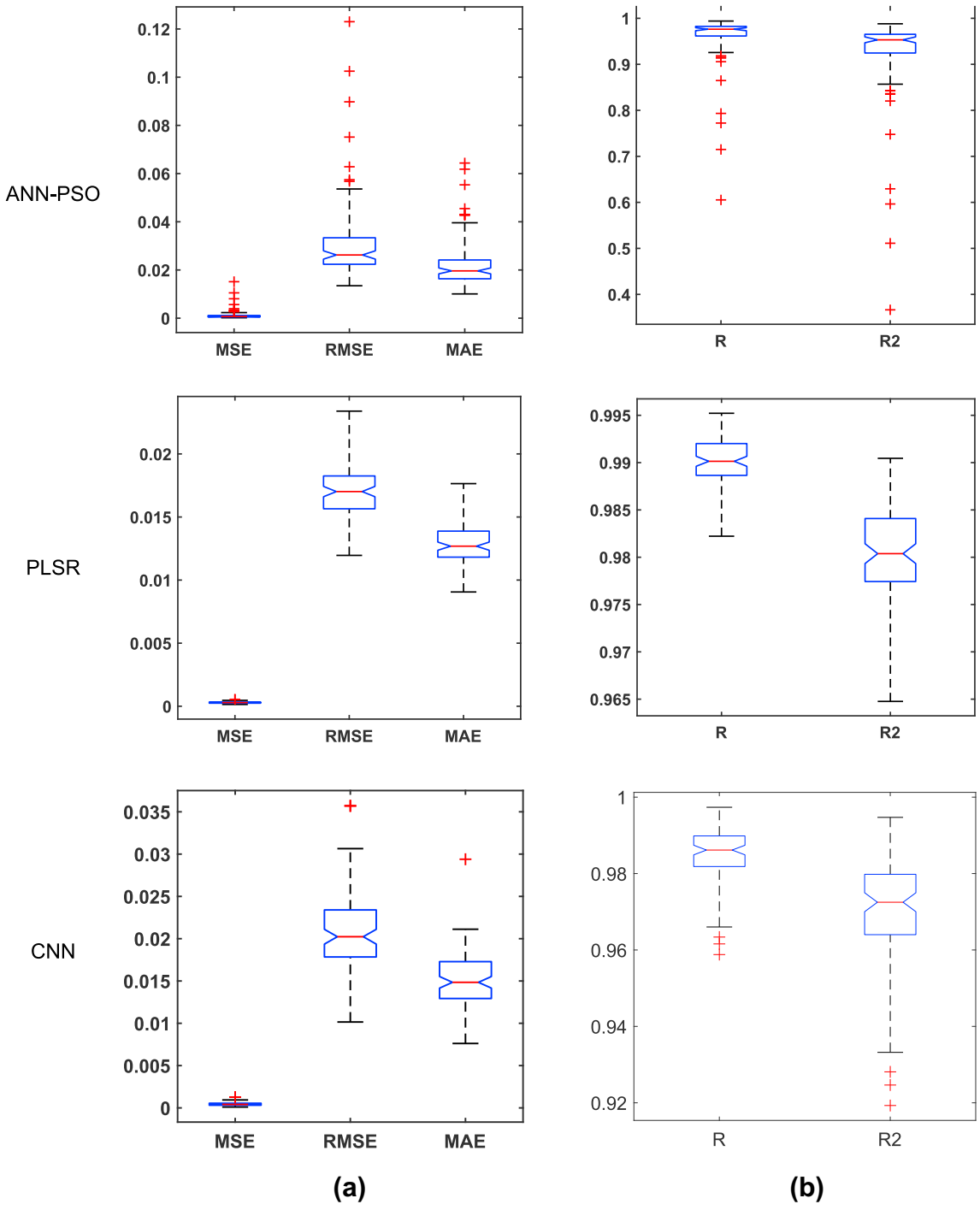

Fig. 3. Boxplots of the five criteria for evaluating the performance of ANN-PSO **(top row)**, PLSR **(middle row)** and CNN **(bottom row)** regression methods for estimating nitrogen content in 30% nitrogen excess treatment in cucumber plant, after 100 iterations (test set): (a) Error criteria (MSE, RMSE and MAE), (b) Regression (R) and determination (R^2) coefficients.

significant difference between the data, each category should be modeled separately. Two statistical test were performed on the extracted spectral images: an ANOVA test to analyze the differences among all the categories, and a Duncan test to analyze the differences between each pair of treatments. The results of the ANOVA test are shown in Table 1, and the results of Duncan test in Table 2. The tests were run with the help of SPSS software (IBM, New York, USA).

According to Tables 1 and 2, the differences between all treatments are statistically significant. This means that we can find differences between the HSI images obtained after the first day of over-dose of nitrogen fertilizer, allowing an early detection of this potential problem. This also supports the idea of generating a separate regression model for each plant category (30%, 60% and 90% nitrogen excess), as described in the following sections.

2.5. Selection of the most effective (optimal) wavelengths for non-destructive estimation of nitrogen

The ultimate goal of the prediction algorithm is to develop a portable predicting device for nitrogen content in plants. Given that the cost and

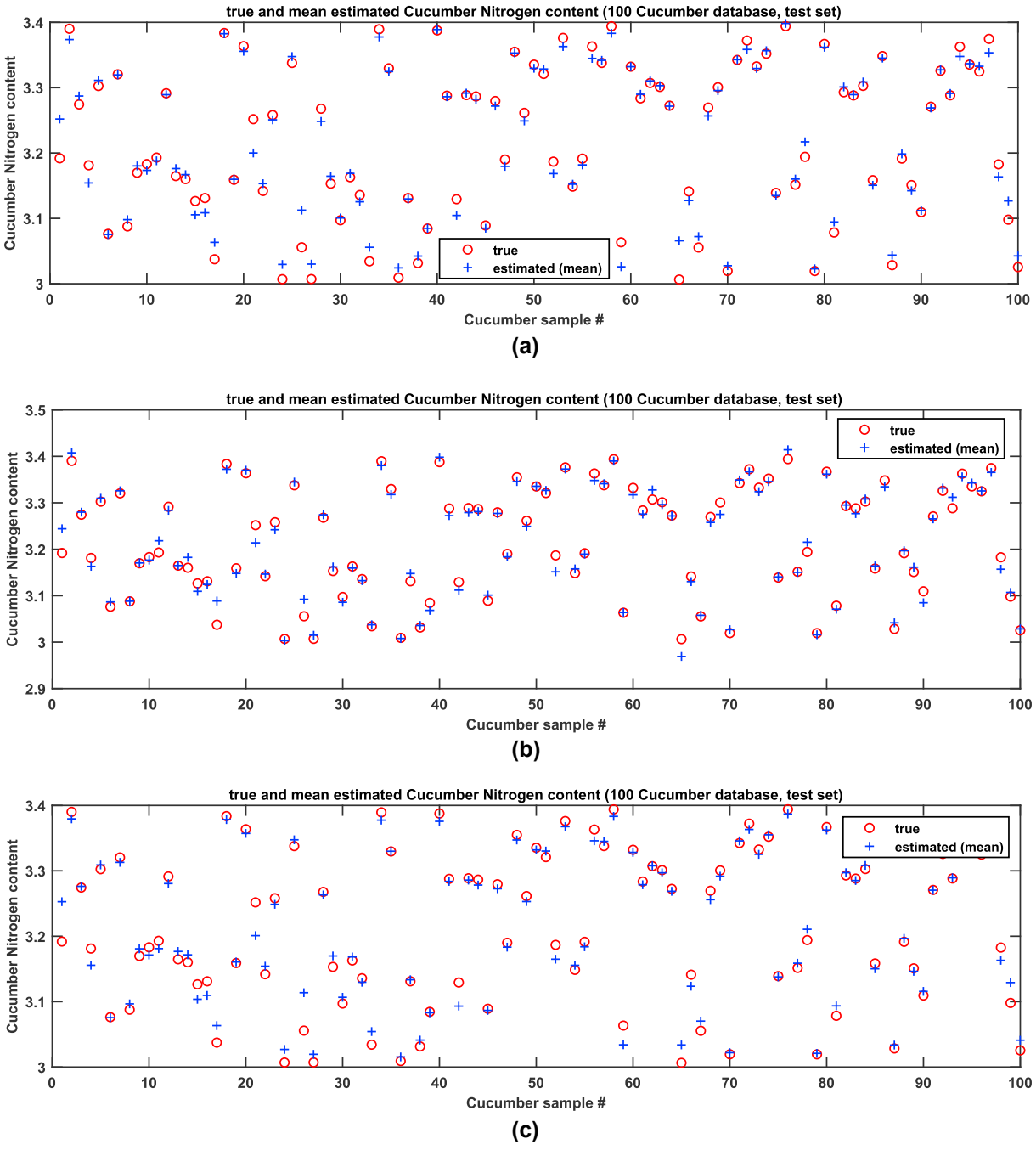

Fig. 4. Measured (true) versus estimated (mean) nitrogen content (mg l^{-1}) in 30% treatment in 100 cucumber plants after 100 iterations (test set). (a) ANN-PSO, (b) PLSR, and (c) CNN.

processing time of portable devices are limiting factors, it is therefore desirable that the number of wavelengths involved in the prediction phase be as low as possible. Therefore, selection of the most effective (discriminant, optimal) spectral wavelengths is important. In this study, a hybrid artificial neural network and simulated annealing (ANN-SA) algorithm was used to select these effective spectral wavelengths.

The SA algorithm is a meta-heuristic algorithm developed by Van Laarhoven and Aarts [26], that solves optimization problems in large input search spaces. The gradual annealing technique is used by metallurgists to achieve a state in which the solid is well organized and its energy is minimized (stable alloy). The aim is to maximize the size of the crystals in the solid state of material. This technique involves placing the material at a high temperature and then gradually lowering its temperature. The founders of this algorithm proposed a method based on gradual and slow annealing technique to solve complex optimization problems, which used computer-based simulation to find the global minimum answer, avoiding being trapped in local minima. In the annealing-based method, each point s in the search space is similar to a state of a physical system, and the function $E(s)$ to be minimized is similar to the internal energy of the system in that state. The goal is to transfer the system from the desired initial state to the state in which the system has least energy possible [27].

In the hybrid ANN-SA approach, the SA method works as a meta-heuristic algorithm controlling the different executions of the ANN, as presented in Fig. S4. The ANN is a regression neural network which takes as input a tuple of wavelength values taken from the pixels of the

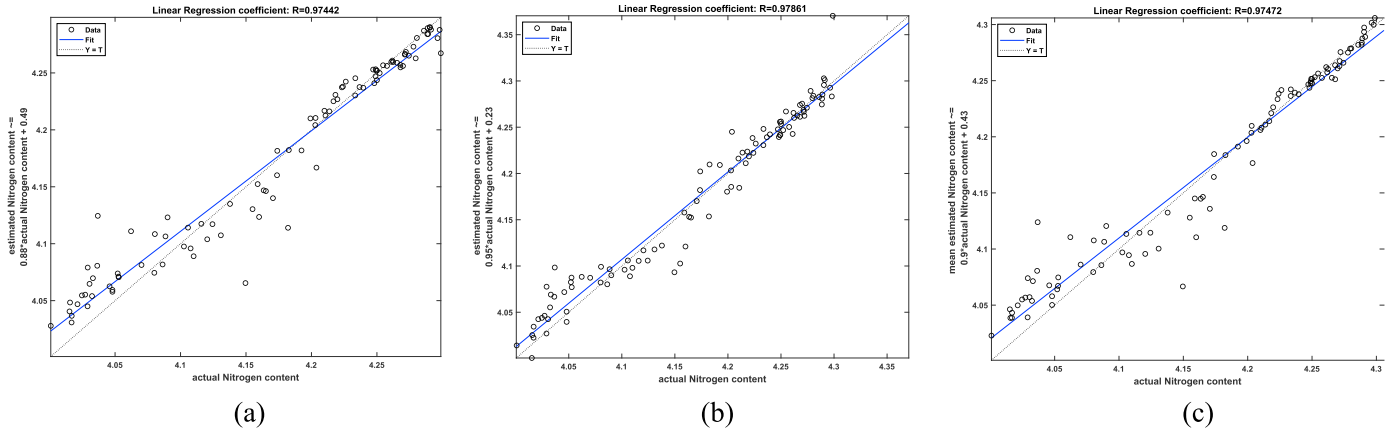

Fig. 5. Scatter plot of regression analysis for mean estimated and measured nitrogen content (mg l^{-1}), in the treatment of nitrogen excess by 60%, after 100 iterations (test set): (a) ANN-PSO, (b) PLSR, and (c) CNN.

hyperspectral images of cucumber leaves, and the expected output is the corresponding nitrogen (N) content values measured by the Kjeldahl method (gold standard), as shown in Fig. S4a. The dataset is decomposed into 70% training, 15% test and 15% validation samples, being all them disjoint sets. The ANN performs a complete training/testing process, obtaining the mean squared error (MSE) over the test set. Resulting accuracy depends on the wavelengths that are used in the input data. For example, if we select wavelengths (600, 800, 1000 nm), this means that the input of the ANN is a tuple of 3 wavelength absorbance (B) values, corresponding to pixels inside HSI images of plant leaves at the given wavelength values. Obviously, background pixels (not belonging to plant leaves) are removed from this process. The structure of the hidden layers of the multilayer (MLP) ANN is given in Table 3.

The selection of the most effective wavelengths is done by the SA algorithm. First, it performs an initial random selection of 3 HSI image absorption wavelength values, and applies the ANN with them. The objective function to minimize is the MSE produced by the ANN, which is computed on the test set. If convergence is not reached, the energy of the system is reduced, and a different selection of wavelengths is done, evolving the current selection according to the current energy of the system. Initially, the variations are larger, and as the system cools down changes get smaller. Then, the ANN is applied again on the new selection of winning wavelengths, and the process is repeated. Finally, the set of wavelengths which produced the least MSE is selected as the optimal set of most effective (discriminant) wavelengths. This process was done in MatLab software (MathWorks, Natick, Massachusetts, USA). More details of the optimization algorithm can be also found in Ref. [27].

2.6. Non-destructive estimation of nitrogen content in cucumber plant leaves

After selecting the most effective wavelengths data with the help of abovementioned ANN-SA method, the next step is to estimate nitrogen (N) content in plant leaves using a regression estimation algorithm. For this purpose, three methods have been implemented and compared next: a hybrid artificial neural network and the particle swarm optimization algorithm (ANN-PSO); partial least squares regression (PLSR); and a convolutional neural network (CNN). These methods are described in the following subsections.

2.6.1. Nitrogen content estimation by an ANN-PSO

PSO is a group-search algorithm that models the social behavior in flocks of birds [28]. The concept of particle refers to a certain solution of the optimization problem. Initially, these particles are randomly located in the solution space. During the execution of the algorithm, particles evolve towards the optimal solution. For doing so, each particle considers its own information and the movement of the neighboring particles. As a result, the swarm of particles tends to move near the best solutions of the problem, according to an objective function to be either maximized or minimized.

In our case, the purpose of the hybrid ANN-PSO approach is to adjust the parameters of the ANN used to estimate nitrogen content in plants. Thus, each particle is a tuple that contains the following parameters: number of hidden layers; number of neurons per hidden layer; transfer function in each layer; back-propagation function; and weight/bias learning function. The number of hidden layers can be selected from 1 to 3, and the maximum number of neurons per hidden layer has been set to 25. The transfer function is selected among 13 different transfer functions available in the neural network toolbox in MatLab; similarly, the back-propagation method is selected among the 19 different functions available, and the weight/bias function among the 15 available methods. So, each particle defines the structure of an ANN, whose input is the tuple of spectral image values at the optimal wavelengths, the output is the estimation of nitrogen content, and the objective function is the MSE, as measured over the test set.

The procedure is similar to the algorithm described for the ANN-SA algorithm. In this way, PSO acts as a metaheuristic algorithm that controls the execution of the ANN for different combinations of the parameters. Fig. S5 presents a diagram of the main steps of the algorithm. For each combination, or particle, the ANN is trained on the train set and tested on the test set. The process begins with a random set of particles, which are evaluated by the ANN; particles move towards the optimal values, until either convergence or a certain number of iterations is reached. Finally, the combination that produced the least MSE value is selected as the optimal configuration of the ANN for the estimation of nitrogen content [29]. A total of 100 replications were executed to evaluate the reliability of the network.

Among a total of 100 input samples, 60% were uniform randomly used as training data, 10% as validation data and the remainder 30% as test data, being all them disjoint sets.

2.6.2. Nitrogen content estimation by linear PLSR

Partial least squares regression (PLSR) method is often called component-based structural equation modeling. It is a linear statistical method proposed by Geladi and Kowalski [30]. Contrary to other methods, such as principal components analysis (PCA), that search for hyperplanes of maximum variance between the output and input variables, PLSR projects the output and input variables into a new space, and then finds a linear regression method in the projected space. PLSR is a non-parametric method, it is robust with respect to the sample size, and does not require any data normalization [24]. Although it was originally applied in econometrics, it has been frequently used also in

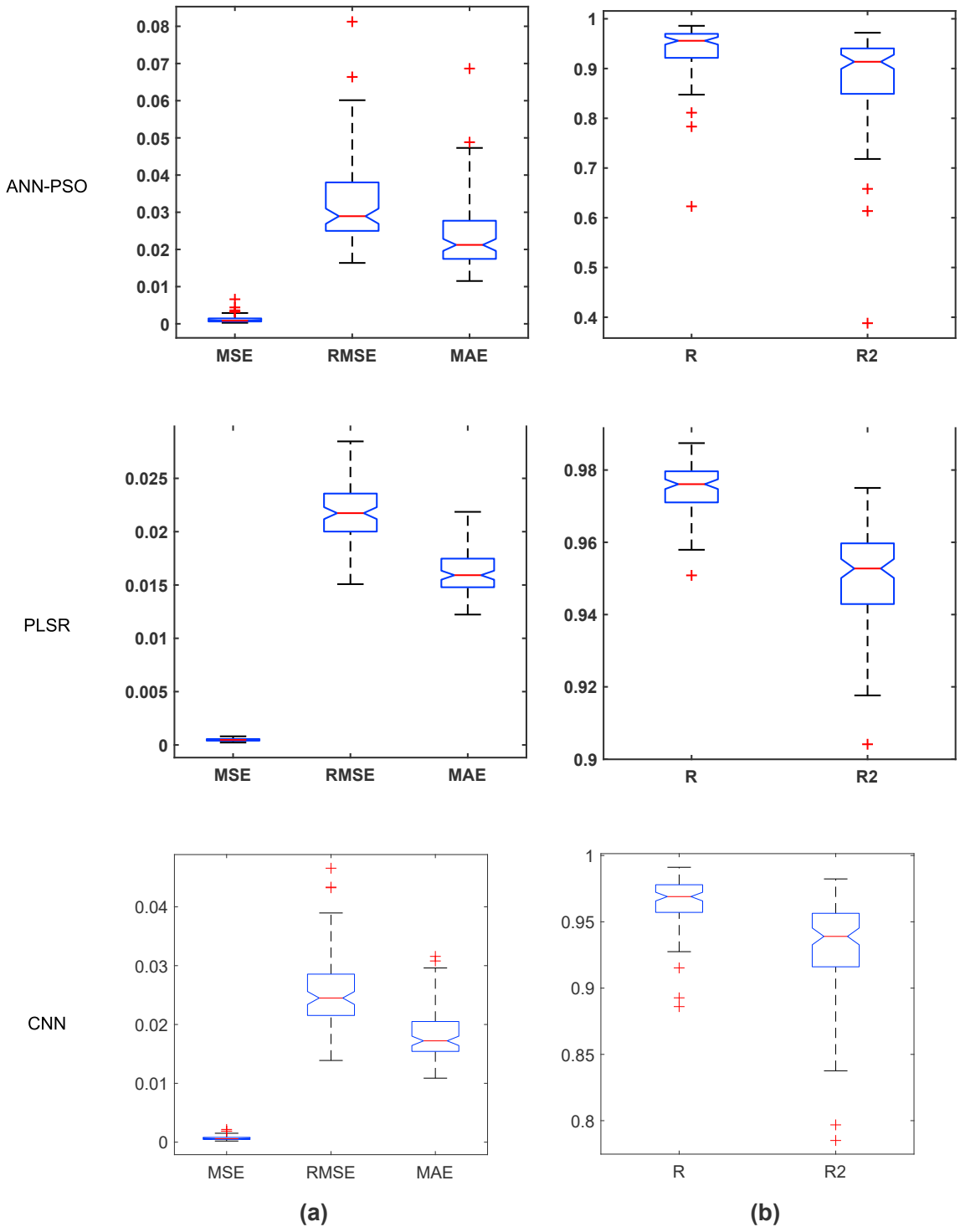

Fig. 6. Boxplots of the five criteria used for evaluating the performance of ANN-PSO **(top row)**, PLSR **(middle row)** and CNN **(bottom row)** regression methods for estimating nitrogen content in 60% nitrogen excess treatment, after 100 iterations (test set): (a) Error criteria (MSE, RMSE and MAE), (b) Regression (R) and determination (R^2) coefficients.

chemometrics, due to its positive properties in this field.

For example, this method is very useful when the effects of several input variables on one or more output variables have to be analyzed. PLSR method can fit the effects of the input, or independent variables, on the output, or dependent variable, as a regression or structural model. If the goal is exploratory predicting or modeling, the application of PLS method is recommended because it does not require any assumptions. It can be applied to a small number of samples and the results are stable against missing information. Technically, the main difference between the PLSR and other regression methods is that the independent variables are placed in several more general factors, so that these factors explain the most of changes in the dependent variables. Further information about PLS regression method can be found in Ref. [31]. For the application of PLSR method to our problem in hand, ParLeS chemometric software was used [24].

While ANN methods (ANN-PSO and CNN) divide the dataset into

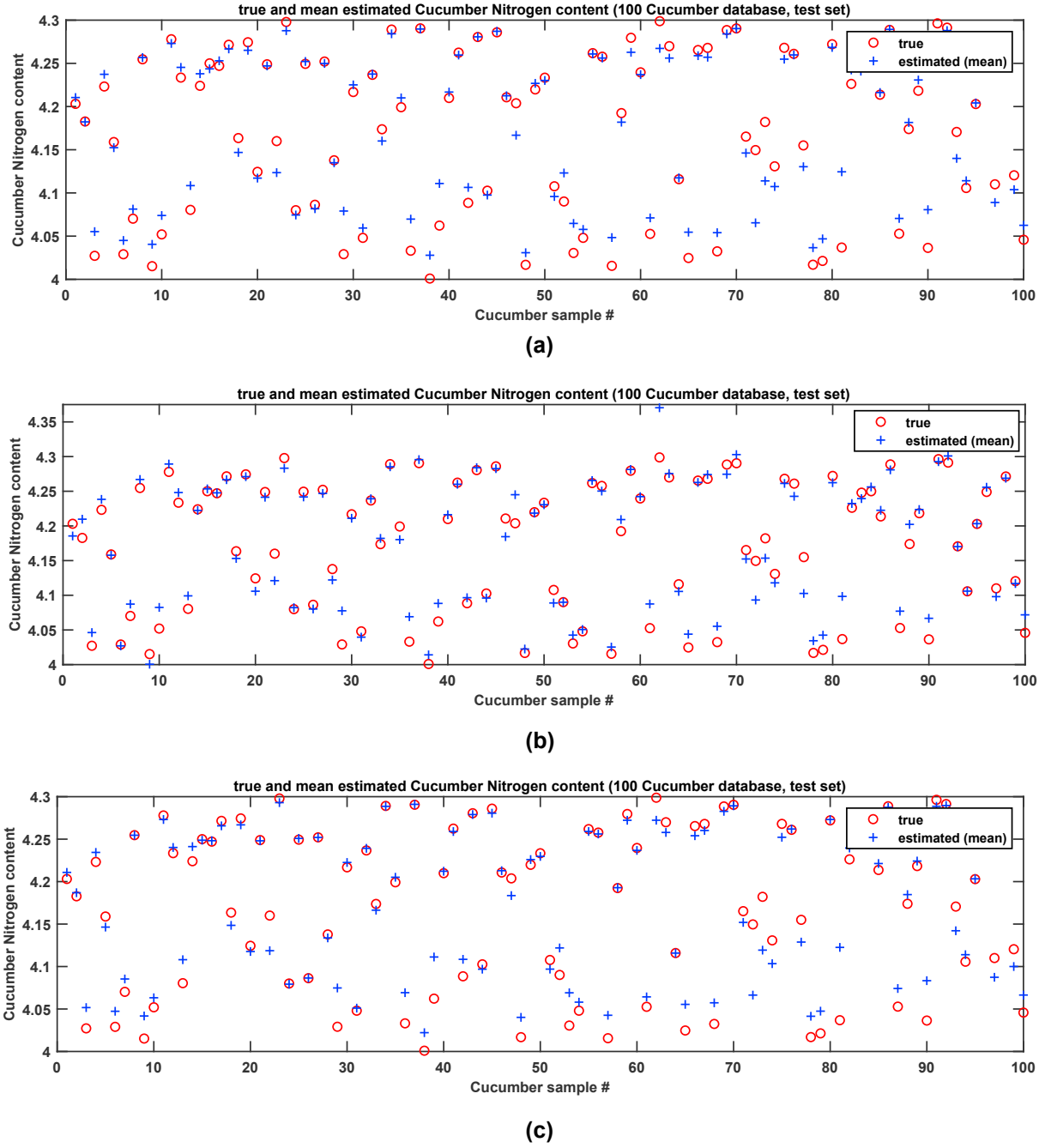

Fig. 7. Measured (true) versus estimated (mean) nitrogen content (mg l^{-1}) in 60% treatment in 100 cucumber plants after 100 iterations (test set): (a) ANN-PSO, (b) PLSR, and (c) CNN.

train/test/validation disjoint subsets, in the case of PLSR only the train and test sets are needed. For consistency in the comparison with the remainder methods, the same partitions were used for all three regressors, including 70% training set and 30% test set in PLSR, being again disjoint sets.

2.6.3. *Nitrogen content estimation by a CNN*

Third proposed regression system for the estimation of nitrogen content in cucumber leaves is based on recent advances on deep learning (DL), and more specifically on convolutional neural networks (CNN) [32], which can still be considered as a part of a more general Machine Learning framework and thus as ANN architectures. As is well-known, the structure of CNNs is inspired by the visual cortex of the human brain. In 1962, Nobel prize winners Hubel and Wiesel, conducted an interesting experiment, observing that edges in different shapes stimulated certain cells in the visual cortex of the brain [33]. CNN network architecture is based in the behavior of the visual cortex of the brain. In fact, on a CNN there are several layers, each specific to identify certain brain items. They are also considered as a type of deep learning algorithm, in the sense that they use neural networks with a large number of hidden layers. These methods require less pre-processing and feature extraction than other classification and regression algorithms, since those steps are done as a part of the network itself.

In our case, the input to the CNN is a 1-dimensional signal given by the 327 wavelength values corresponding to each pixel of hyperspectral plant leaf images. Observe that, in this method, the regressor is not limited to the 3 most effective wavelengths, as in ANN-PSO and PLSR. A typical structure of CNN consists of different layers of convolutions, with

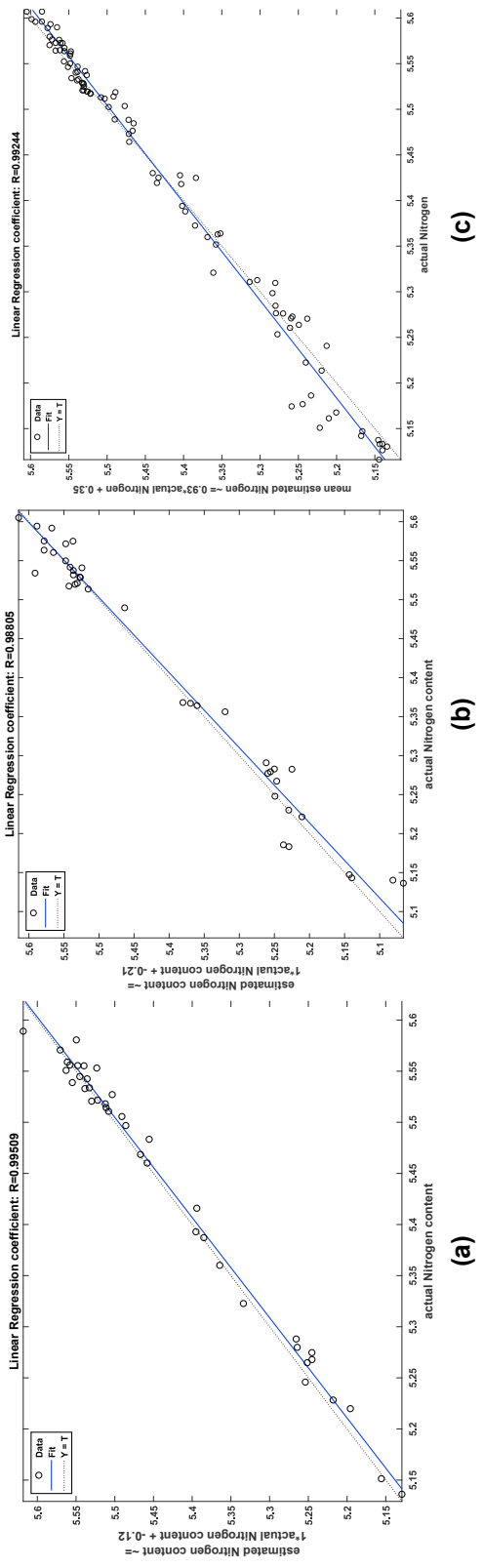

Fig. 8. Scatter plot of regression analysis for mean estimated and measured nitrogen content($mg \cdot l^{-1}$) in the treatment of nitrogen excess by 90%, after 100 iterations (test set): (a) ANN-PSO, (b) PLSR, and (c) CNN.

ReLu (rectified linear) non-linear activation function, and max pooling, finishing with a dense layer, as depicted in Fig. S6. The output is a single neuron that estimates nitrogen content value for the input image sample.

After a process of trial and error of different structures for the CNN, changing the number of layers, filters and convolution sizes, the selected configuration is shown in Table 4. The selected setup contains 4 steps of convolution + ReLu + max pooling, until reducing the width to only 17 values of 256 features, an additional convolutional layer, a flatten layer, and a final dense layer to produce the desired regression result. For consistency, the same disjoint partition of train, validation and test sets as the previous method was done, although there are two main differences: (i) the input tuple contains all the 327 wavelengths, as already mentioned; and (ii) data augmentation was used to have more virtual samples to perform the training process, since the number of available samples is very limited for an effective application of deep learning methodology. These virtual samples were obtained as weighted averages of the real samples inside the training set. This model was implemented in Python using Keras deep learning framework [34].

2.7. Evaluation of the performance of the regression methods

In order to evaluate the performance of the three proposed methods for predicting nitrogen content in cucumber plant, the following well-known criteria were used: coefficient of determination (R^2), regression coefficient (R), mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE), as detailed in Refs. [35,36]. These criteria measure the difference between nitrogen content estimated by non-destructive regressors from hyperspectral images and true (golden standard) nitrogen content as measured by destructive Kjeldahl chemical method.

3. Results and discussion

In this section, the experimental results of the proposed models are presented and discussed. First, results of the ANN-SA method for selecting the most effective wavelengths are described. Then, the regression models for each N class treatment (30%, 60% and 90% excess nitrogen) are shown. Finally, the three proposed models based on ANN-PSO, PLSR and CNN, are numerically compared, over the test set.

3.1. Selection of the most effective spectral wavelengths imaging data

As described, the first step of the research process consists of extracting the most effective imaging wavelengths data by a hybrid ANN-SA algorithm. In all cases, the system selected top three wavelength numbers. For 30% nitrogen excess treatment, the selected wavelengths were 812, 924 and 987 nm. For 60% treatment, they were 799, 879 and 903 nm. And for 90% nitrogen excess 883, 896 and 949 nm. It is interesting to observe that all selected wavelengths are located in the NIR range (750–1000 nm), even though the available data also included the visible range (400–750 nm). This indicates that a simple computer vision methodology based on standard RGB visible range images could not be adequate to properly estimate nitrogen content in plant leaves, since other parts of the spectrum are needed. This can also be observed in Fig. 1, where some sample hyperspectral cucumber plant leaf images are shown.

As depicted in Fig. 1, the differences from treatment to treatment are more evident at the selected optimal wavelengths, showing a greater reflectance as nitrogen (N) content increases. It can also be observed that the degradation of plant leaves by nitrogen over-dose is more noticeable in the tips of cucumber plant leaves.

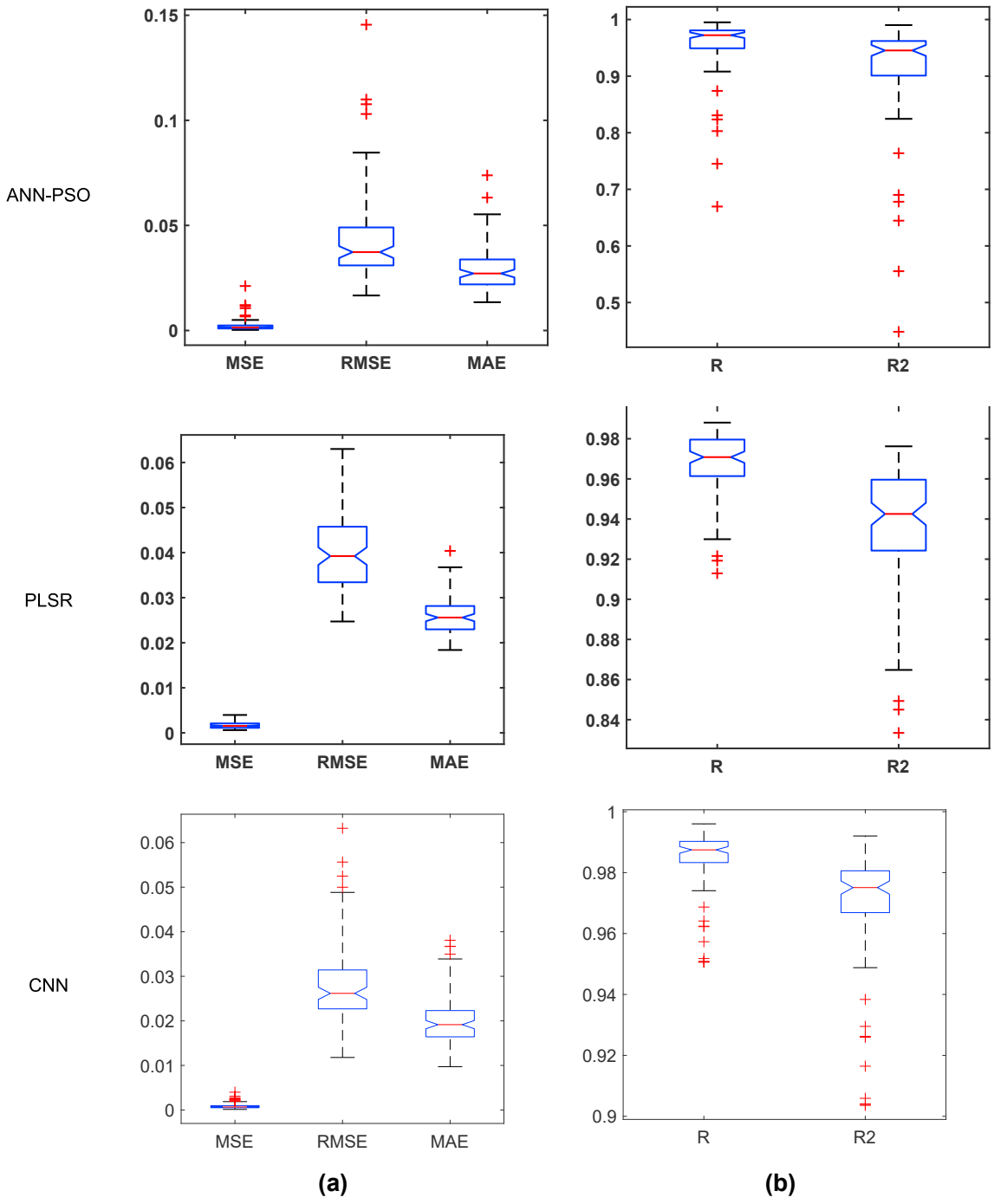

Fig. 9. Boxplots of the five criteria for evaluating the performance of ANN-PSO **(top row)**, PLSR **(middle row)** and CNN **(bottom row)** regression methods for estimating nitrogen content in 90% nitrogen excess treatment, after 100 iterations (test set): (a) Error criteria (MSE, RMSE and MAE), (b) Regression (R) and determination (R^2) coefficients.

3.2. Performance of proposed regression models: ANN-PSO, PLSR and CNN

3.2.1. Application of nitrogen content excess by a 30%

To examine the reliability of the predicting algorithms, a total of 100 iterations were executed for each model, choosing uniform random train, validation and test disjoint sets input samples. Fig. 2 shows the scatter plot of regression analysis for mean estimated and actual nitrogen in cucumber plant based on spectral data of effective wavelengths in 100 iterations of the process. The values of the linear correlation coefficient for predicting nitrogen content were 0.991 for ANN-PSO, 0.992 for PLSR, and 0.992 for CNN, indicating that all regression models were able to predict nitrogen content in plant leaves with high accuracy.

Fig. 3 shows the boxplot of the criteria used for evaluating the performance of the regression methods after 100 iterations, including regression (R) and determination coefficient (R^2) boxplots. Fig. 4 depicts measured and mean estimated nitrogen content values for each of the 100 input samples after 100 iterations, for ANN-PSO (Fig. 4a), PLSR (Fig. 4b) and CNN (Fig. 4c) models. As seen, estimated nitrogen content mean values are almost always very close to (true) measured values, for all the 100 input samples.

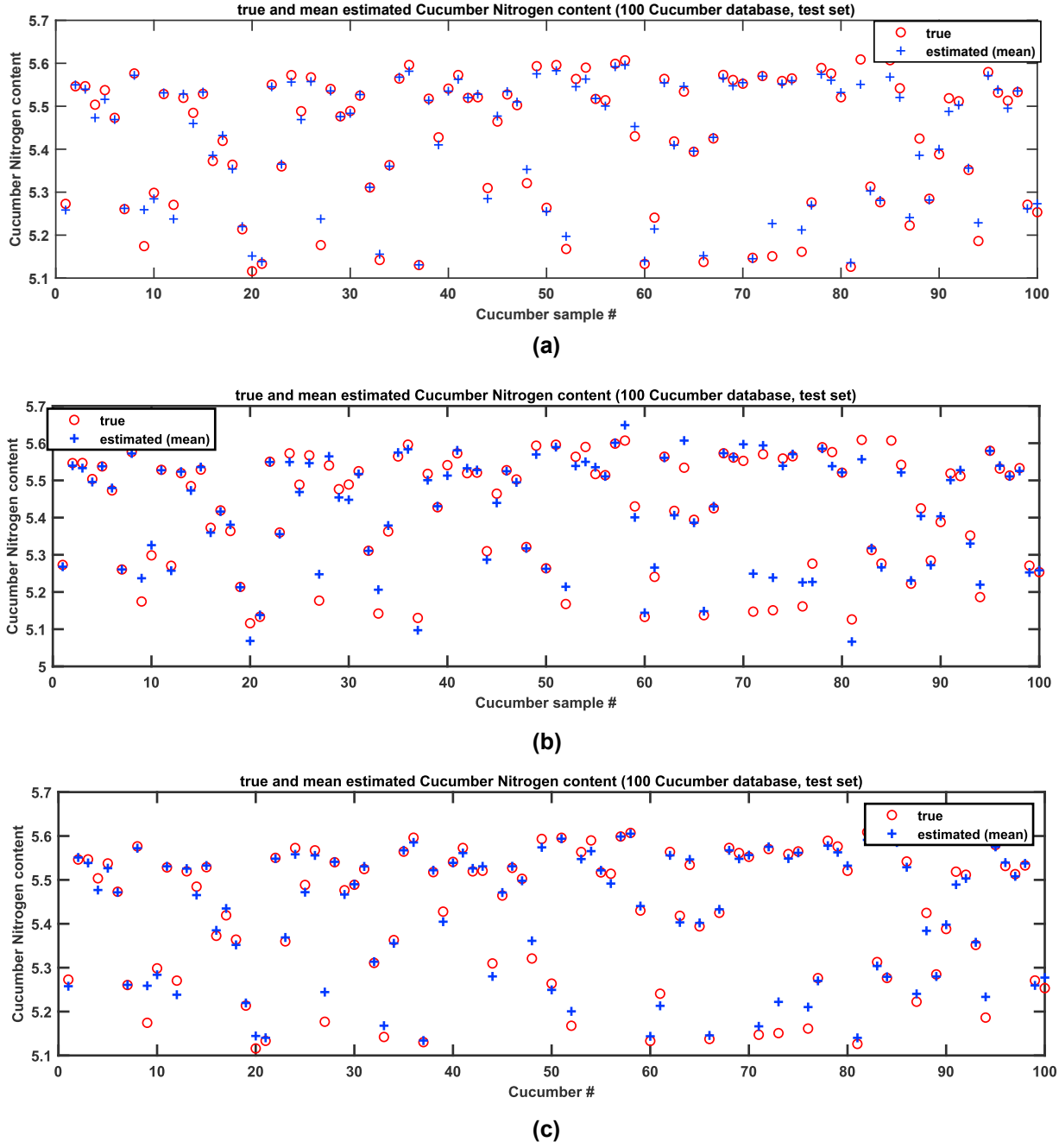

Fig. 10. Measured (true) versus estimated (mean) nitrogen content (mg l^{-1}) in 90% treatment in 100 cucumber plants after 100 iterations (test set): (a) ANN-PSO, (b) PLSR, and (c) CNN.

3.2.2. Application of nitrogen content excess by a 60%

Fig. 5 shows the scatter plot of regression between mean estimated and measured amount of nitrogen content in cucumber plants for the case of nitrogen excess by 60%. The linear correlation coefficients were 0.974 for ANN-PSO, 0.979 for PLSR and 0.975 for CNN, indicating the high accuracy of all the methods. Fig. 6 shows the boxplots of the performance criteria after 100 iterations, including regression (R) and determination coefficient (R^2). Fig. 7 presents a comparison of measured (true) and mean estimated nitrogen content values. Again, high accuracy can be observed in the estimated values for most of the input samples. It can be observed that the highest errors are normally produced for the lowest values of nitrogen content, while the average and highest concentrations produce less error.

3.2.3. Application of nitrogen content excess by a 90%

As in the previous treatments, results for the 90% nitrogen excess are presented in three figures. Fig. 8 shows the scatter plots of the mean estimated and measured (true) nitrogen values; Fig. 9 the boxplots of the performance criteria after the 100 iterations; and Fig. 10 the comparison of measured and estimated mean nitrogen content values. In this case, the correlation coefficients were 0.995 for ANN-PSO, 0.988 for PLSR and 0.992 for CNN. So, although all the methods are able to achieve remarkable good results, PLSR is not the best method, and it is slightly overcome by both ANN-PSO and CNN.

Table 5
Comparison of mean ± standard deviation for various performance criteria in the estimation of nitrogen content in cucumber plants after 100 iterations using ANN-PSO, PLSR and CNN regression methods (test set): MSE: mean squared error; RMSE: root mean squared error; MAE: mean absolute error; R: regression coefficient; R^2: determination coefficient.

	N excess (%)	MSE	RMSE	MAE	R	R^2
ANN-PSO	30	0.0012 ± 0.0020	0.0312 ± 0.0171	0.0222 ± 0.0100	0.960 ± 0.055	0.925 ± 0.093
	60	0.0011 ± 0.0009	0.0324 ± 0.0113	0.0237 ± 0.0080	0.937 ± 0.051	0.882 ± 0.088
	90	0.0022 ± 0.0028	0.0431 ± 0.0200	0.0292 ± 0.0100	0.965 ± 0.049	0.926 ± 0.086
PLSR	30	0.0003 ± 0.0000	0.0171 ± 0.0021	0.0119 ± 0.0030	0.990 ± 0.002	0.980 ± 0.001
	60	0.0004 ± 0.0001	0.0218 ± 0.0026	0.0160 ± 0.0040	0.975 ± 0.006	0.951 ± 0.012
	90	0.0016 ± 0.0007	0.0400 ± 0.0087	0.0257 ± 0.0040	0.997 ± 0.016	0.986 ± 0.003
CNN	30	0.0004 ± 0.0002	0.0208 ± 0.0046	0.0150 ± 0.0032	0.985 ± 0.007	0.970 ± 0.013
	60	0.0006 ± 0.0003	0.0256 ± 0.0061	0.0183 ± 0.0043	0.965 ± 0.018	0.933 ± 0.034
	90	0.0009 ± 0.0006	0.0284 ± 0.0090	0.0205 ± 0.0058	0.984 ± 0.010	0.969 ± 0.019

3.3. Comparison of the performance of ANN-PSO, PLSR and CNN regression methods for estimating nitrogen content in cucumber plant leaves using effective wavelengths data

After analyzing in the previous sections the results achieved by each model, Table 5 presents the comparison of their performance criteria, including: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), regression coefficient (R) and coefficient of determination (R^2). For each method (ANN-PSO, PLSR and CNN) and treatment (30%, 60% and 90% excess) the mean and standard deviation values of the 100 iterations are computed and shown.

It can be observed that all the proposed methods are able to achieve remarkable high accuracies, with R mean values always between 0.937 and 0.997 range, and mean MSE below 0.0022. This can be considered as a very promising result compared to other works in the literature. A direct comparison with previous research cannot be done, since the problem, objectives and dataset of our study are not the same as theirs. However, we can observe that the typical regression coefficients (R) for estimating different properties of the plants using spectroscopy are around values of 0.85–0.96 for the method by Serranti et al. [16], from 0.885 to 0.993 for Eshkabilov et al. [18], and from 0.957 to 0.971 for Wang et al. [20]. Thus, the accuracy of our method is inside those of the state of the art using similar spectroscopy techniques, although in a different scenario.

On average, PLSR obtains the best results for all three nitrogen treatments, closely followed by CNN. These two methods show a very similar accuracy, while ANN-PSO has a slightly worse performance. In addition, PLSR has consistently the best R value for all the cases. At the same time, CNN is able to overcome error values of PLSR for the 90% nitrogen content treatment, with a mean MAE of only 0.0205, against 0.0257 of PLSR.

Although CNN could be expected to be superior to other techniques, this is not the case in our problem. We believe that some specific characteristics of our case of study differ from other classical problems where DL and CNN have often proven to overcome other existing methods. First, the spectral information is given by a tuple of 327 wavelengths (o just 3 values, if we consider only the most effective ones), while typical applications of CNN are done on samples with high dimensionality (for example, images or videos). In fact, it would be nonsense to apply CNN to input tuples of just 3 values. Second, as already mentioned, the available input dataset is quite reduced (100 times 3, 300 total cucumber plant leaves). Although we have applied data augmentation, it was proven not to be able to generalize in an effective way to the test set. We could try to extend the dataset with more samples, but since the experiment requires costly laboratory analysis, it would be difficult to obtain datasets that reach some thousand samples, which is common in most DL applications.

Another very positive aspect of PLSR is that the standard deviations of the performance criteria are very low. For example, in the 90% treatment, it has a mean R of 0.997 and standard deviation of 0.016, which is the largest standard deviation value for PLSR approach. These reduced values indicate that the model is robust and stable under different scenarios and iterations. CNN method also presents this positive characteristic, while ANN-PSO has a lower stability and exhibits higher values of the standard deviation. Finally, we can remark that although the 90% excess treatment is the case that presents the lowest error in all three models, the 30% category is also estimated with a very high precision, e.g. a mean R of 0.990 in PLSR and 0.985 in CNN. This is a very positive sign, since it indicates that the proposed methodology could be applied to detect nitrogen (N) over-use in the early stages, when the misuse of fertilization could be corrected by farmer on time.

4. Conclusion

Abusive application of chemical fertilizers overturns the ecological balance, eventually destroying many organisms and causing potential problems and hazards for the human body. Reliable diagnosis of the nutritional status of agricultural products is an essential part in farm management, being plant leaves an important source of information for identifying them. Therefore, the development of intelligent systems can help non-destructive and real-time portable detection of plants content, like the one here proposed that need few input data (pixel and wavelength light absorbance values) from plant to accurately estimate N content values. So far, several methods have been developed for non-destructive prediction in agricultural products, some of which have been justified in terms of real interest applicability. Hyperspectral imaging (HSI) has been extensively studied and developed by integrating both spectroscopic and imaging techniques. For this reason, the purpose of our work was to study the feasibility of using HSI as a non-destructive method to detect and avoid in an early stage excessive use of nitrogen fertilizers. The study has been focused on cucumber plants, and three regression techniques have been proposed and compared, based on a hybrid artificial neural networks and the particle swarm algorithm (ANN-PSO), partial least squares regression (PLSR) and convolutional neural networks (CNN). The first two use a reduced set of the 3 most effective wavelengths, while the third one uses all the wavelengths data available.

All the proposed regression algorithms were able to accurately predict nitrogen content in cucumber plants using spectral data, being PLSR slightly ahead of CNN and ANN-PSO, both in MSE and R mean values. Since the estimation accuracy is very high even for the lowest level of over-application of N fertilizer (30%), this approach can be effectively used to detect small deficiencies in the fertilization of the plants, before the contamination of the plants gets unrecoverable. In any case, more research would be necessary to extend these results to other types of plants. For example, it was observed that the systems tend to produce larger error for the lowest levels of nitrogen content. Besides, although CNN showed promising results, it is evident that considerably larger datasets are required to make them work to their full potential. Another future line of research is to achieve a unified model for prediction of nitrogen content, instead of requiring one model for each type of treatment.

CRediT authorship contribution statement

Sajad Sabzi: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Formal analysis, Investigation, Data curation, Visualization. **Razieh Pourdarbani:** Resources, Project administration, Funding acquisition, Writing – review & editing. **Mohammad H. Rohban:** Resources, Writing – review & editing. **Ginés García-Mateos:** Methodology, Validation, Investigation, Data curation, Visualization, Supervision, Writing – original draft, Writing – review & editing. **Juan I. Arribas:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Visualization, Supervision, Project administration, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded in part by the Spanish Ministry for Science, Innovation and Universities (MICINN), Agencia Estatal de Investigación (AEI), as well as by the Fondo Europeo de Desarrollo Regional funds (FEDER, EU), under grant numbers RTI2018-098958-B-I00 (J.I. Arribas) and RTI2018-098156-B-C53 (G. Garcia-Mateos).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2021.104404.

References

[1] Y. Zhai, X. Zhao, Y. Teng, X. Li, J. Zhang, J. Wu, R. Zuo, Groundwater nitrate pollution and human health risk assessment by using HHRA model in an agricultural area, NE China, Ecotoxicol. Environ. Saf. 137 (2017) 130–142.
[2] Y.M. Liu, D.Y. Liu, W. Zhang, X.X. Chen, Q.Y. Zhao, X.P. Chen, C.Q. Zou, Health risk assessment of heavy metals (Zn, Cu, Cd, Pb, as and Cr) in wheat grain receiving repeated Zn fertilizers, Environ. Pollut. 257 (2020) 113581.
[3] N.S. Bryan, H. van Grinsven, The role of nitrate in human health, Adv. Agron. 119 (2013) 153–182.
[4] H.J. Buscaglia, J.J. Varco, Early detection of cotton leaf nitrogen status using leaf reflectance, J. Plant Nutr. 25 (9) (2002), 2067–2067.
[5] A. Sabu, K. Sreekumar, Literature review of image features and classifiers used in leaf based plant recognition through image analysis approach, in: 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 145–149), IEEE, 2017.
[6] M. Agarwal, S. Gupta, K.K. Biswas, A new Conv2D model with modified ReLU activation function for identification of disease type and severity in cucumber plant, Sustain. Comput.: Inf. Syst. (2020a) 100473.
[7] S. Sabzi, R. Pourdarbani, J.I. Arribas, A computer vision system for the automatic classification of five varieties of tree leaf images, Computers 9 (1) (2020) 6.
[8] M. Agarwal, S. Gupta, K.K. Biswas, Development of Efficient CNN model for Tomato crop disease identification, Sustain. Comput.: Inf. Syst. 28 (2020b) 100407.
[9] D.A. Dean, T. Ramanathan, D. Machado, R. Sundararajan, Electrical impedance spectroscopy study of biological tissue, J. Electrost. 66 (3–4) (2008) 165–177.
[10] M. Salimi, R. Pourdarbani, B. Asgarnezhad Nouri, Factors affecting the adoption of agricultural automation using Davis's acceptance model (case study: Ardabil), Acta Technol. Agric. 23 (1) (2020) 31–39.
[11] N. Nesakumar, C. Baskar, S. Kesavan, Analysis of moisture content in beetroot using fourier transform infrared spectroscopy and by principal component analysis, Sci. Rep. 8 (2018) 7996.
[12] J.A. Kim, D.J. Wales, G.Z. Yang, Optical spectroscopy for in vivo medical diagnosis—a review of the state of the art and future perspectives, Prog. Biomed. Eng. 2 (4) (2020), 042001.
[13] H. Yuping, L. Renfu, C. Kunjie, Prediction of firmness parameters of tomatoes by portable visible and near-infrared spectroscopy, J. Food Eng. 222 (2018) 185–198.
[14] s. Jarolmasjed, R. Lav, S. Sindhuja, Hyperspectral imaging and spectrometry-derived spectral features for bitter pit detection in storage apples, Sensors 18 (2018) 1561.
[15] L. Hee, Ch Byoung-Kwan, Rapid assessment of corn seed viability using short wave infrared line-scan hyperspectral imaging and chemometrics, Sensor. Actuator. S0925–4005 (17) (2017), 31458–2.
[16] S. Serranti, A. Trella, G. Bonifazi, C. Garcia Izquierdo, Rapid monitoring of physical-chemical parameters by hyperspectral imaging, Waste Manag. 75 (2018) 141–148.
[17] S. Zhou, L. Sun, W. Xing, G. Feng, Y. Ji, J. Yang, Sh Liu, Hyperspectral imaging of beet seed germination prediction, Infrared Phys. Technol. 108 (2020) 103363.
[18] S. Eshkabilov, A. Lee, X. Sun, C.W. Lee, H. Simsek, Hyperspectral imaging techniques for rapid detection of nutrient content of hydroponically grown lettuce cultivars, Comput. Electron. Agric. 181 (2021) 105968.
[19] Sh Chen, T. Hu, L. Luo, Q. He, Sh Zhang, M. Li, X. Cui, H. Li, Rapid estimation of leaf nitrogen content in apple-trees based on canopy hyperspectral reflectance using multivariate methods, Infrared Phys. Technol. 111 (2020) 103542.
[20] Y.J. Wang, G. Jin, L.Q. Li, Y. Liu, Y. Kianpoor, J.M. Ning, Z. Zhang, NIR hyperspectral imaging coupled with chemometrics for nondestructive assessment of phosphorus and potassium contents in tea leaves, Infrared Phys. Technol. 108 (2020) 103365.
[21] J. Sun, W. Yang, M. Zhang, M. Feng, L. Xiao, G. Ding, Estimation of water content in corn leaves using hyperspectral data based on fractional order Savitzky-Golay derivation coupled with wavelength selection, Comput. Electron. Agric. 182 (2021) 105989.
[22] N. Liu, F.A. Townsend, M.R. Naber, P.C. Bethke, W.B. Hills, Y. Wang, Hyperspectral imagery to monitor crop nutrient status within and across growing seasons, Rem. Sens. Environ. 255 (2021) 112303.
[23] E. Amoah, Z. Du, Y. Lu, Y. Hu, Detection and assessment of nitrogen effect on cold tolerance for tea by hyperspectral reflectance with PLSR, PCR, and LM models, Inf. Process. Agric. 8 (1) (2020) 96–104.
[24] R.A.V. Rossel, ParLeS: software for chemometric analysis of spectroscopic data, Chemometr. Intell. Lab. Syst. 90 (1) (2008) 72–83.
[25] R.B. Bradstreet, Kjeldahl method for organic nitrogen, Anal. Chem. 26 (1) (1954) 185–187.
[26] P.J. Van Laarhoven, E.H. Aarts, Simulated annealing, in: Simulated Annealing: Theory and Applications, vols. 7–15, Springer, Dordrecht, 1987.
[27] A. Zameer, S.M. Mirza, N.M. Mirza, Core loading pattern optimization of a typical two-loop 300 MWe PWR using Simulated Annealing (SA), novel crossover Genetic Algorithms (GA) and hybrid GA(SA) schemes, Ann. Nucl. Energy 65 (2014) 122–131.
[28] D. Wang, D. Tan, L. Liu, Particle swarm optimization algorithm: an overview, Soft Computing 22 (2) (2018) 387–408.
[29] S. Sabzi, Y. Abbaspour-Gilandeh, G. García-Mateos, A new approach for visual identification of orange varieties using neural networks and metaheuristic algorithms, Inf. Process. Agric. 5 (1) (2018) 162–172.
[30] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[31] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130.
[32] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B.,. Shuai, T. Chen, Recent advances in convolutional neural networks, Pattern Recogn. 77 (2018) 354–377.
[33] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106–154.
[34] A. Gulli, S. Pal, Deep Learning with Keras, Packt Publishing Ltd, 2017.
[35] R. Pourdarbani, S. Sabzi, D. Kalantari, R. Karimzadeh, E. Ilbeygi, J.I. Arribas, Automatic non-destructive video estimation of maturation levels in Fuji apple (Malus Malus pumila) fruit in orchard based on colour (Vis) and spectral (NIR) data, Biosyst. Eng. 195 (2020) 136–151.
[36] M. Alibaba, R. Pourdarbani, M.H. Khoshgoftar Manesh, G.V. Ochoa, J.D. Forero, Thermodynamic, exergo-economic and exergo-environmental analysis of hybrid geothermal-solar power plant based on ORC cycle using emergy concept, Heliyon 6 (4) (2020), e03758.