Sustainable Computing: Informatics and Systems 28 (2020) 100415

Contents lists available at ScienceDirect

Sustainable Computing: Informatics and Systems

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

A self-adaptive classification method for plant disease detection using GMDH-Logistic model

Junde Chen[a], Huayi Yin[b,c,d], Defu Zhang[a,*]

[a] School of Informatics, Xiamen University, Xiamen, 361005, China
[b] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, Fujian, 361024, China
[c] Post Doctoral Station of Computer Science and Technology, School of Informatics, Xiamen University, Xiamen, 361005, China
[d] Postdoctoral Workstation of Jinjiang High Tech Park, Jinjiang, 362200, China

ARTICLE INFO

Article history:
Received 14 January 2019
Received in revised form
17 November 2019
Accepted 25 May 2020
Available online 26 June 2020

Keywords:
Plant disease detection
Classification method
Index system
GMDH-Logistic model

ABSTRACT

Agriculture occupies an important position in the economic development, and which is one of the important sources of earning for human beings in many countries. However, a number of challenges are faced by the farmers including different diseases for plants, as having diseases in plants are quite natural. The primary premise in the prevention and treatment of diseases is to correctly identify and judge the disease during the growth of plants, so as to further grasp the rules of disease development. As a matter of fact, an automated plant disease detection technology can be more beneficial for monitoring the plants, and the leaves of plant are the first source of plant disease detecting, the most diseases may be detected from the symptoms appeared on the leaves. Therefore, this paper introduces a novel approach for the automatic detection and classification of plant leaf diseases. Based on the image process techniques, we perform the feature engineering analysis and build the index system for the prediction models. Then the selected features are fed to the GMDH-Logistic model, and the comparison experiments are performed. The outcomes illustrate that the method is effective, it can identify whether the plant is the diseased plant or not.

1. Introduction

The occurrence of plant diseases can have significant negative effects on the quality and quantity of agricultural products, and if the plant diseases are not detected in time, there will be an increase in food insecurity [1]. Therefore, the early warning and forecast is the basis of effective prevention and control for the plant diseases, and it plays an important role in the management and decision-making for agricultural production. The traditional identification of plant diseases are mainly based on the visual observations of experienced producers or plant experts in the field, this method is time-consuming, labor-intensive, lags in prediction, and it cannot be carried out in a wide range [2,3], it is impossible to perform the real-time and rapid identification for plant diseases. However, the rapid development of image processing technology has provided a new way for the disease recognition [4–7], with the development of image processing and pattern recognition technologies, it provides a scientific basis for the identification of plant diseases based on leaf images, therefore, this manuscript proposes a novel strategy to identify the plant diseases, and which not only improves the recognition efficiency, but also solves problems such as lack of agriculture experts and poor objectivity.

Generally, plant leaves are the first source to detect most of the plant diseases. Plant diseases can be detected automatically through efficient image processing techniques [8,9]. However, the identification of plant diseases through image processing is not an easy job because of the huge disparities available in the leaves of different and similar plants for image noise, instance size, color, texture, shape, etc. Various image processing strategies have been anticipated to overcome such problems and normally all methods have two steps [8]. In the first phases prominent features are extracted from the input images of the leaves, many researchers proposed different feature extraction methods based on data mining such as intra and inter-block dependencies [10] for Markov features, for spatial domain subtractive pixel adjacency model (SPAM) [11], bag of visual words [12], convolutional neural network [13] and many more. These methods automatically generate high dimensional features without human experts. However, high dimensionality [14] is a major concern in the case of images. At present, there are two common approaches of data dimensionality

* Corresponding author.
E-mail address: dfzhang@xmu.edu.cn (D. Zhang).

https://doi.org/10.1016/j.suscom.2020.100415
2210-5379/

reduction: feature extraction and feature selection. Among them, principal component analysis (PCA) [15] and clonal selection algorithm [16] are two classical methods and can be effectively utilized for dimension reduction, while the challenge is the interpret of extracted variables and the optimal parameter values, additionally, information distortion is also a problem.

In the second phase, a particular classifier is used which classifies the images into healthy or diseased images. Major classification techniques that are popularly used for disease identification in plant are k-nearest neighbor (kNN) [17], support vector machine (SVM) [18], fisher linear discriminant (FLD) [19], artificial neural network (ANN) [20], random forest (RF) [21] and so on. For example, Phadikar et al. [22] used two approaches to classify rice diseases, including the Bayes and SVM classifiers. In their work, 10 diff ;erent combinations of training and test data resulted in an accuracy of 79.5 %. And the SVM achieved 68.1 % accuracy with 10-fold cross-validation. Sandeep Kumar et al. [23] used SVM classifier which classified normal and diseased potato leaf into two diff ;erent classes, and their classifier achieved an accuracy of 92.12 %. Kahar et al. [24] used the ANN as a classifier to perform the disease detecting of rice, in which 3 output nodes were used for detecting three diseases: leaf blight, leaf blast, and sheath blight. Their identification scheme achieved an accuracy of 74.21 %, etc.

More recently, deep learning techniques, particularly convolutional neural networks (CNN), has shown outstanding performance in image processing and classification [25–27]. So far, the application of deep learning methods in agriculture is not much [28], most of the existing studies focused on laboratory scenarios with simple background conditions. Mohanty, Hughes, and Marcel [29] trained a deep learning model for recognizing 14 crop species and 26 crop diseases. Their trained model achieves an accuracy of 99.35 % on a held-out test set. H. Gensheng et al. [27] used a deep convolutional neural network for tea disease detection; they realized a recognition accuracy of 92.5 %. Kawasaki et al. [30] introduced a system based on CNN to recognize cucumber leaf disease; it achieves an accuracy of 94.9 %, etc. Although very good results have been reported in the literature, most investigation materials include solely images in experimental (laboratory) setups, not in real field scenarios. In reality, plant leaf images are captured under heterogeneous conditions and include an extensive variety of symptom characteristics [31]. The effects of those data constraints were illustrated by Mohanty et al., who observed a quick drop in accuracy when the model trained on the laboratory database was applied to images collected online [32].

Therefore, referring to the above literature researches, this paper introduces a novel identification method of plant diseases, based on the leaf images, we perform the feature engineering analysis and build the index system for the prediction models, thus the selected features are fed to the models and the comparison experiments are performed as also. The experimental outcomes illustrate that the model is effective, and it can identify whether the plant is the diseased plant or not.

2. Analysis process and methods

2.1. Preprocessing

In the process of image formation, the image quality may be degraded due to the imperfection of the imaging system, transmission medium and recording equipment, which can cause the poor visual effects and the difficulties in computer processing. That is, the factors including the imaging system, environment, noise, imaging characteristics, etc., may affect the image quality. So, it is necessary to preprocess the image before image analysis, and in the project, the preprocessing work such as gray transformation and image denoising was performed in advance.

2.1.1. Gray transformation

The uneven distribution of illumination conditions and the different time or position for photographing, which may cause the non-uniform gray scale of images, we can assume that the plant leaf image $I(x, y)$ is composed of gray background $I_b(x, y)$, leaf diseases $I_n(x, y)$ and random noise $I_c(x, y)$ in the image, then this could be expressed as the following Eq. (1).

$$I(x, y) = I_b(x, y) + I_h(x, y) + I_b(x, y) \tag{1}$$

Therefore, the extraction of background images is critical, and the bilinear interpolation method is used in the gray transformation of images, thus the background images are extracted, and the main process is listed as follows.

(1) Obtain the background subset. The image is divided into blocks, and the background point is taken from each block. The uneven illumination may be obvious in the whole image, but it can be considered as nearly uniform in the local area.
(2) Implement the bilinear interpolation. Since the gray transformation of image is a gradual process, the bilinear interpolation method can be used to interpolate the four adjacent pixels, so that the generated surface is continuous.

The image of gray transformation is as follows (Fig. 1).

2.1.2. Image denoising

Just as mentioned in the previous section, the system is inevitably affected by external factors, which can cause some noise in the image and has a negative impact on feature extraction. Therefore, the image denoising should be performed, and the classical filtering approaches such as median filtering and Gaussian filtering are commonly used, for the long history of development, they have been widely used and proved to be correct and superior in some problems too.

The median filtering is to calculate the median of all pixels in the template and use the calculated median to change the gray value of a center pixel in the template. This method is not so sensitive to noise and can eliminate salt and pepper noise better, but it also easily leads to the discontinuity of the image. While suppressing the noise, it has relatively large damage to the image edges and details, which makes the image edges and details blurred, so, on the basis of this, a gradient inverse weighted approach is taken and compared with the median method.

For a discrete image, the grayscale change in a region is less than that between regions, and the absolute value of the gradient at the edge is higher than that inside the region. In an $n \times n$ window, if the inverse of absolute gradient value between the central pixel and its neighbors is defined as the weight of each neighbor point, the weight value in the region is the largest, while the weight value in the edge and outside is smaller. Therefore, instead of the arithmetic average, the weighted average could improve the algorithm in theory and not blur the edges too much. The detailed process is as follows.

Set the gray value of the point (x, y) is $f(x, y)$, then in the 3*3 neighborhood of this point, the gradient inverse can be computed in Eq. (2).

$$g(x, y; i, j) = \frac{1}{f(x+i, y+j) - f(x, y)} \tag{2}$$

where $i, j = -1, 0, 1$, but i and j are not equal to 0 at the same time. Thus the gradient inverse of 8 neighbor points can be calculated, in particular, if $f(x+i,y+j) = f(x, y)$, the value of $g(x, y; i, j)$ is defined in

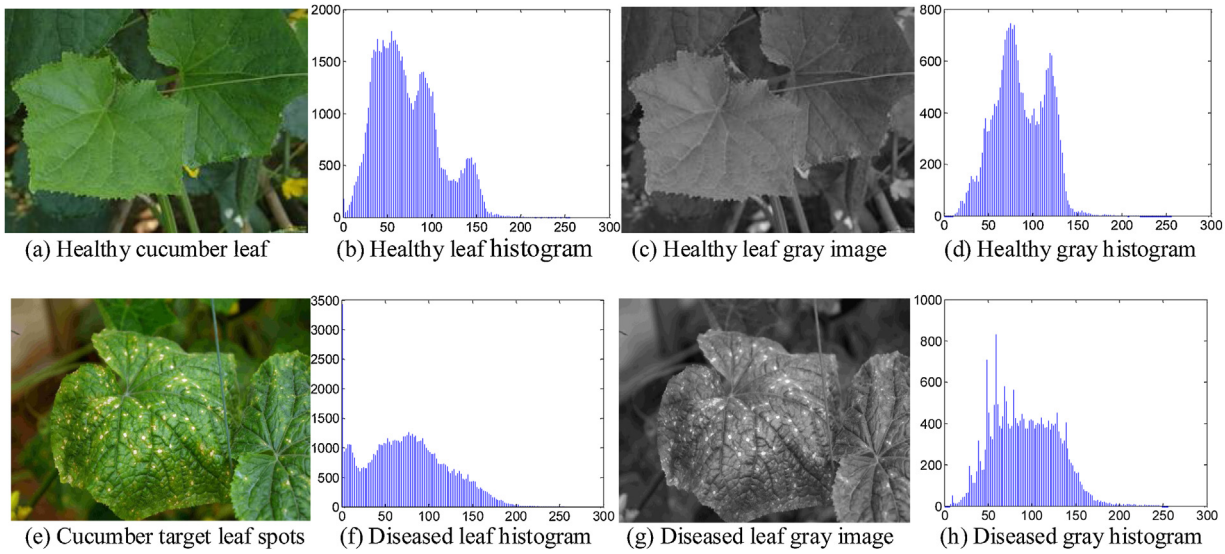

Fig. 1. The examples of gray transformation for cucumber plant leaf images.

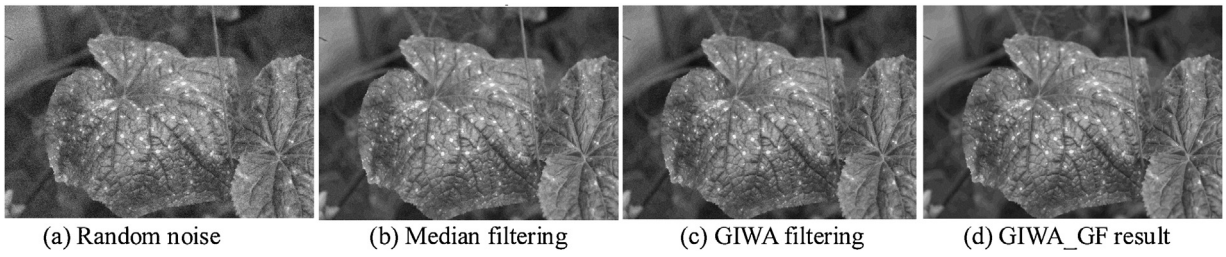

Fig. 2. The filtering algorithm comparison.

the interval [0, 2], so, a normalized weight matrix template can be obtained as follows.

$$W = \begin{bmatrix} w(x-1,y-1) & w(x-1,y) & w(x-1,y+1) \\ w(x,y-1) & w(x,y) & w(x,y+1) \\ w(x+1,y-1) & w(x+1,y) & w(x+1,y+1) \end{bmatrix} \tag{3}$$

The weight of the central point is set as $w(x, y) = 0.5$, and the sum weights for the other 8 points is 0.5, so that the sum weights of all points is 1. Thus, for each point, the weight can be computed in Eq. (4).

$$w(x+i,y+j) = \frac{1}{2}\frac{g(x,y;i,j)}{\sum_i \sum_j g(x,y;i,j)} \tag{4}$$

where $i, j = -1, 0, 1$, but i and j are not equal to 0 at the same time.

It can be seen from Fig. 2(c) that the gradient inverse weighted algorithm(GIWA) has a certain effect for the image noise, which can smooth some of the noise information in the background and retain the edge information of the diseases. Especially for the serious diseases, this algorithm has a good effect to some extent, but the filtering of this algorithm is not very obvious, the noise is not eliminated completely. So, the method of GIWA followed by Gaussian filtering (GIWA_GF) is adopted in the paper, after the GIWA filtering, the Gaussian filtering is implemented again, thus some outlier points can be removed fully, and the edge details of leaf diseases are well preserved as well, as shown in Fig. 2(d).

2.2. Image segmentation

Image segmentation partitions an image into non-overlapping regions, which is an essential step in computer vision, image analysis and image understanding [33,34]. The segmentation of leaf image is to separate the leaf disease areas from its background and form a binary image for the following feature extraction and computing. Since the colors of different types of leaves are not the same, it is difficult to segment them with a uniform gray threshold, therefore, the gray threshold must be calculated for each image.

The gray levels in the same region are often similar, while the gray levels in different regions show significant differences. When the gray level difference between two regions separated by threshold t is large, the difference between the average gray μ_1, μ_2 of the two regions and the total average gray μ of the image is also large. So, the regional gray variance can express this characteristic, and the formula is as follows.

$$\sigma_B^2 = \delta_1(t)[\mu_1(t) - \mu]^2 + \delta_2(t)[\mu_2(t) - \mu]^2 \tag{5}$$

where σ_B^2 is the regional grey variance, and $\delta_1(t)$, $\delta_2(t)$ are the area ratio of the segmented regions to the whole image. Obviously, with different threshold t, the different variance would be obtained, that is, the regional average gray, area ratio and variance are all the function of threshold t, after that, the total average gray μ of the image can be computed as

$$\mu = \mu_1(t)\delta_1(t) + \mu_2(t)\delta_2(t) \tag{6}$$

Then, the σ_B^2 could be computed in Eq. (7).

$$\sigma_B^2 = \delta_1(t)\delta_2(t)[\mu_1(t) - \mu_2(t)]^2 \tag{7}$$

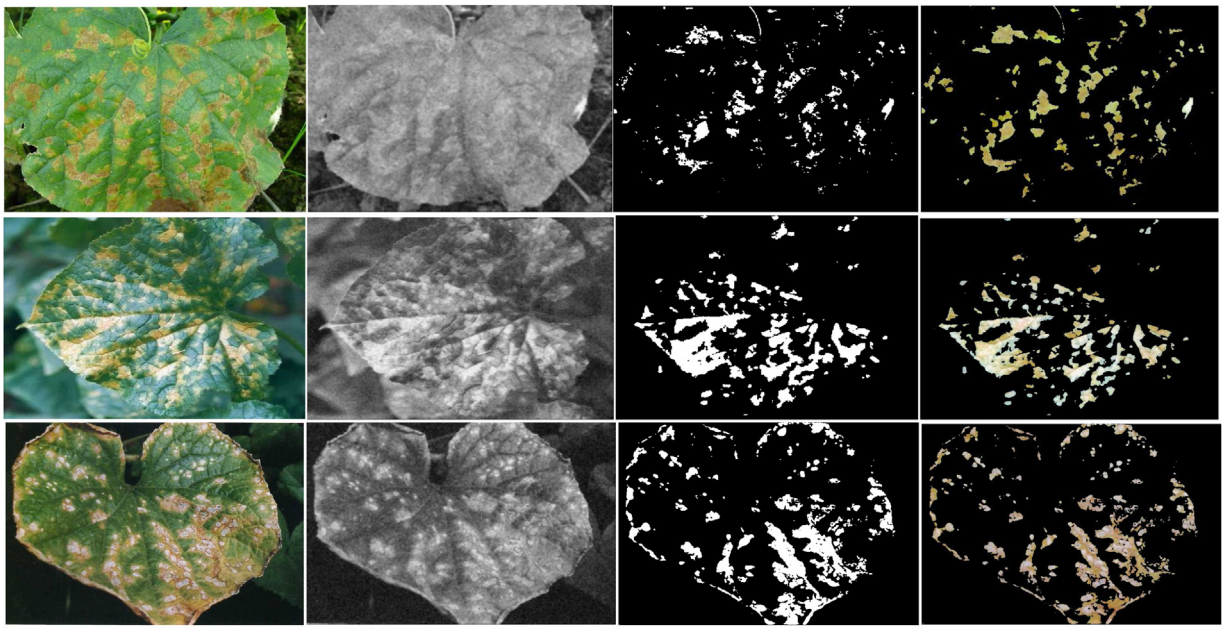

Fig. 3. The image segmentation of cucumber leaf lesions.

when the variance value between two segmented regions reaches the maximum, it is considered to be the optimal segmentation state of the two regions, thus the threshold t is determined. Therefore, based on this analysis, the normalized gray value could be used as the threshold to binarize the image accordingly. The partial segmentation images are as follows.

As can be seen in Fig. 3, the disease spot regions of plant leaf images are basically extracted out and no much background information is retained in the segmented images, so this segmented images can be used for the next feature extraction and further to perform the classification of cucumber plant disease images.

3. Feature extraction

In the process of image analysis, the specific and unique characteristics in the image are often interested. By separating the useful information for classification and identification, the image features could be formed and extracted, on the basis of this, further computing could be carried out. The image feature extraction refers to extract the key indicators that it can reflect the essence of the images, so that the purpose of automatic recognition or evaluation for the images could be achieved. Actually, feature extraction is one of the critical steps in image recognition or classification, and the result of which directly affects the quality of the model. For the leaf images, there are many features that can be extracted, such as the geometric features, texture features, color features, statistical features, etc., this paper focuses on extracting the leaf image features from the above aspects.

3.1. Leaf image features

3.1.1. Texture feature extraction

The texture refers to the recurring patterns and rules in images, the texture of the normal and diseased leaf images is very different in thickness and direction. Therefore, the characteristics of smoothness, roughness, and granularity for the disease spots can be quantified by measuring the change of the gray scale. The gray level co-occurrence matrix (GLCM) of the image can reflect the comprehensive information of image gray for direction, adjacent interval, variation amplitude, etc., so it is often used as the basis to quantitatively describe the attributes of texture features. The following four variables are computed from the GLCM and used as the texture feature indicators.

(1) Contrast. The contrast is computed as $contrast = \sum_{i,j} |i-j|^2 p(i,j)$, and the range of which is [0 (*I*-1)2], where $p(i, j)$ is the value of the *i-th* row and *j-th* column in the gray level co-occurrence matrix, I is the total row number of gray level co-occurrence matrix.
(2) Energy. The energy is computed as $energy = \sum p(i,j)^2$, which is the sum of the squares of all elements in the gray level co-occurrence matrix, the range of value is [0 1].
(3) Correlation. The correlation is computed as $correlation = \sum_{i,j} \frac{(i-\mu_1)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j}$, where μ, σ are the mean and standard deviation respectively. The range of value is [-1 1].
(4) Homogeneity. The homogeneity is computed as $homogeneity \sum_{i,j} \frac{p(i,j)}{1+|i-j|}$, and the range of which is also the [0 1].

3.1.2. Color feature

Compared with the RGB mode, HSI mode including hue, saturation, and intensity is more consistent with the human eye's perception of color, and it can separate the color information from the brightness information of the image, which enables the computer to transform the recognition pattern according to different illumination. However, the symptom information of diseases in a single color space is very limited, so, in addition to the RGB features, the HSI features are extracted at the same time, thus the comprehensive color space features including the HSI and RGB are adopted.

3.1.3. Disease area feature

The various degrees of the disease show different sizes and shapes in the leaves, which make it possible to describe the sever-

Table 1
The extracted feature data (partial).

Grades	Feature (s_1)		Feature (s_2)					Feature (s_3)							
	pct	num	r	g	b	h	s	ct1	ct2	cr1	cr2	en1	en2	hm1	hm2
slight	0.19	23189	4093131	4796258	3685760	6639	2859	2.69	2.28	0.72	0.76	0.59	0.60	0.90	0.91
slight	0.06	5082	951905	961374	564329	856	1646	0.54	0.58	0.81	0.79	0.87	0.87	0.98	0.98
normal	0.08	20677	4613992	4721099	4658507	10241	616	0.28	0.16	0.95	0.97	0.85	0.85	0.99	0.99
moderate	0.11	35471	6729118	7016603	6084723	9481	3574	1.10	1.15	0.82	0.81	0.75	0.75	0.95	0.95
normal	0.00	8	1437	1672	1068	2	2	0.00	0.00	0.26	0.24	1.00	1.00	1.00	1.00
normal	0.14	19073	3515654	3664591	2543740	3754	4213	0.95	0.74	0.86	0.89	0.72	0.72	0.96	0.97
normal	0.03	2006	362633	442572	333155	615	360	0.93	0.71	0.58	0.67	0.91	0.92	0.98	0.98
serious	0.17	20487	4164938	4582450	3787165	5390	2076	1.31	1.27	0.88	0.88	0.66	0.66	0.95	0.95
normal	0.01	828	148854	135249	49088	123	508	0.07	0.06	0.74	0.77	0.98	0.98	1.00	1.00
slight	0.02	7550	1346840	1634923	960111	1863	2160	0.57	0.57	0.65	0.64	0.93	0.94	0.98	0.98
normal	0.01	1892	345443	383972	275878	420	348	0.27	0.25	0.70	0.72	0.96	0.96	0.99	0.99
slight	0.00	20677	4613992	4721099	4,658,507	6639	2859	0.28	0.16	0.95	0.97	0.85	0.85	0.99	0.99

ity of the disease by the area percentage of disease spots. For the morphologies of the disease, spots are only reflecting the shape features, it is not necessary to consider the color attribute of leaf diseases. Here, we can convert the gray images of leaves into the binary images, and which not only makes the operation simple but also saves the storage space. Since the value of the binary image is either Öör ẗ255, it is marked with 0 or 1 respectively. So, for the area percentage of disease spots, we can have

$$Pct = \frac{\sum P_{ij}}{\sum P} \tag{8}$$

where P_{ij} is the pixel in disease areas, P is the pixel in total image.

3.1.4. Disease pixel feature

The pixel values of different disease images are all not the same, for a binary image of a blade, we can define the pixel values as the following Eq. (9).

$$Num = \sum_{i=1}^{N}\sum_{j=1}^{M} I(i,j) \tag{9}$$

where Num is the sum of pixel values in a binary image, and $I(i, j)$ is the value of one pixel point in the image. From the above formula, the pixel values of various types of diseases can be calculated. The pixel values of normal images are the smallest, while that of serious disease images is the largest. So, since the pixel values are different for various disease images, we could define the pixel values as a feature of the disease types, and use it to reflect the classification information of leaf diseases.

3.2. The extracted indicators

As analyzed in the previous section, there are three major categories, and 15 feature indicators are extracted, they are the morphological feature class S_1, color feature class S_2 and texture feature class S_3 separately. Among them, the morphological features are mainly extracted from the disease-spots area percentage and the sum of pixel values, which is S_1 ={*Pct, Num*}. The color features include both RGB color features and HIS color features, and the values of R, G, B, H, S are extracted respectively, that is S_2 ={R, G, B, H, S}. The texture features are mainly extracted from the comprehensive information about the direction, adjacent interval and variation amplitude of image grayscale reflected by the GLCM, it mainly includes the indicators of contrast, correlation, energy, and homogeneity, that is S_3 ={$Contrast_1$,$Contrast_2$,$Correlation_1$,$Correlation_2$,$Energy_1$, $Energy_2$, $Homogeneity_1$, $Homogeneity_2$}. The extracted feature value data of plant leave images are listed in Table 1.

As shown in Table 1, after the feature indicators are extracted from the leaf images, the feature selection should be performed, and the feature variables that have a critical impact on the target and high contribution to the classification need to be selected to ensure more effective recognition and lower model complexity.

4. Classification and detection

4.1. Model building

The principal component analysis (PCA) is a classical method for reducing the dimensionality of data and has been widely investigated in pattern analysis especially image recognition, it is based on the variable covariance matrix to process, compress and extract features, so as to reduce the computational complexity and assist in different recognition tasks, however, this method does not consider the relationship between the independent variable and the dependent variable, and when the explanatory variable is too much, the extracted principal component is generally difficult to interpret [35,36]. In addition, the feature selection methods based on optimization calculations such as genetic algorithm [37], particle swarm optimization [38], etc., the optimization algorithms often contain many parameters, and their optimal values are difficult to be determined too.

The GMDH (Group Method of Data Handling) [39] method provides an idea for this issue, which can automatically determine the variables, structures, and parameters entering the model. It overcomes the disadvantages of genetic algorithm(GA) to a certain extent, at the same time, the features selected by the GMDH method are usually interpretable and can make up for the shortcomings of PCA. The network structure of GMDH is shown in Fig. 4.

The classical GMDH algorithm uses the linear function as its transfer function, and the result of which is the detailed numerical data, it is often used to handle the data prediction issue. However, the plant disease detection in the real business is often complicated and linear inseparability, in this situation, it is unreasonable to use the linear transfer function to represent the relationship between response variables and characteristic variables. Therefore, this paper combined the advantages of the Logistic and GMDH algorithms to propose a new recognition algorithm GMDH-Logistic, the main improvement points are listed as follows.

4.1.1. The transfer function

In the network layer, all possible pairs of the m inputs are generated to create the transfer functions of the $k = m^*(m\text{-}1)/2$ neurons and the traditional GMDH linear function is replaced by the non-

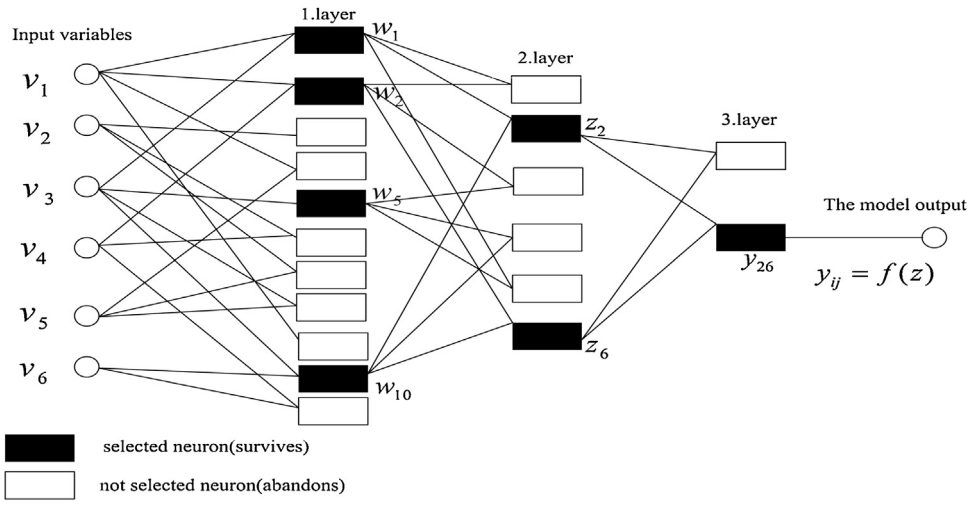

Fig. 4. The GMDH network structure.

linear function, here, the logistic function is used, and the formula is given as follows.

$$y_{ij} = \frac{e^{a_0+a_i v_i+a_j v_j}}{1+e^{a_0+a_i v_i+a_j v_j}}. \tag{10}$$

where y is the response variable, it represents the predicted classification of the model. The v_i, v_j were the selected competition modeln the previous network, which is combined pairs for the next layer input. a_0, a_i, a_j. a_0, a_i, a_j is the coefficient vector of the model.

4.1.2. External criterion

The traditional GMDH method is often used the accuracy criteria as the external criterion, the main representation of accuracy criteria is the regularity criteria. But as for the plant disease detection, it could be correctly identified whether the plant is diseased or not, which is more than the detection accuracy. That is to say, the principle of recall is re important than the precision principle in this issue. Hence, the recall rate criteria rather than the accuracy criteria is used as the external criterion in the GMDH-Logistic model, and which is computed using Eq. (11).

$$Recall\ Rate = \frac{TP}{TP+FN} \times 100\% \tag{11}$$

where the *TP* (True positive) is the number of instances that actually belong to the class *C* and are correctly identified by the classifier, *FN* (false negative) is on the contrary, which is the number of instances that belong to the class *C* but mistakenly classified.

The details of enhanced algorithm was listed as above Fig. 5, so, on the basis of this, we could use the GMDH-Logistic to perform the identification of plant diseases, and as mentioned in the previous section, the commonly used methods such as the artificial neural networks, support vector machine and other optimization algorithms, etc. were applied for the comparison analysis at the same time.

4.2. Model detection

In the paper, there are 80 cucumber leaf images of normal, minor, moderate, and serious disasters, and each category includes 20 images, the illumination of the images does not change much and the background conditions are complex, all the size of the images has been normalized. Therefore, we can divide the images into two groups according to the ratio of 4:1, and one is for the training, the other is for the testing. The detailed steps of GMDH-Logistic are listed as follows.

1 Divide the given dataset *D* into the training set *A* and testing set *B*, then *D* = *A*+ *B*. The method of *k*-fold cross-validation is used (*k* = 5), and the dataset is equally divided into *k* parts at random.
2 The formula(10) is used as the new transfer function, and the formula(11) is used as the external criterion. The cross training set A is applied to train the GMDH-Logistic model.
3 Through combining every two nodes of the former layer, the candidate models are generated to get the new input nodes. The unknown parameters were estimated on the training set by inner criterion (least squares).
4 The selected intermediate models survived, and they were used as the inputs for the next layer to create the new generation of models while the non-selected models are abandoned.
5 Thus, steps 2–4 are repeated until the optimal complexity model is obtained.
6 The testing set B is applied to validate, and logistic function was used for the final classification output of the samples.

So, on the basis of the above steps, we could perform the identification of plant diseases by the GMDH-Logistic model. Just as mentioned in the previous section, the algorithm can automatically determine the variables, structures, and parameters entering the model. Take the first test for example, there are 5 indicators selected as the input variables, which are Pct, Contrast1, Contrast2, Energy1, Homogeneity1 respectively, and the prediction result is shown as follows.

As can be seen in Table 2, except for 2 leaf images of slight disease and moderate disease, the other 14 leaf images with diseases are basically detected out, therefore, the recall rate of diseased plants is 85.71 %, in the same way, the other four parts are used as the testing set separately, total of 5 tests were performed. Each time, the training dataset is used to learn the model and the testing dataset is used to test the model, the 5 different results were obtained respectively. At the same time, the other commonly used algorithms such as the artificial neural networks, support vector machine, etc. are selected for the comparative analysis in the testing, where the PCA-SVM is to perform the PCA first and then implement the SVM algorithm, the PCA-ANN is similar to PCA-SVM, and GANN refers to the genetic artificial neural networks. Particularly, the convolutional neural networks (CNN), which is a state-of-the-art machine learning model for image recognition, is also selected in our comparative analysis.

Considering the non-massive labeled images of the dataset, the new samples are generated to enrich the dataset using data augmentation techniques. By random rotation, flipping, and scale

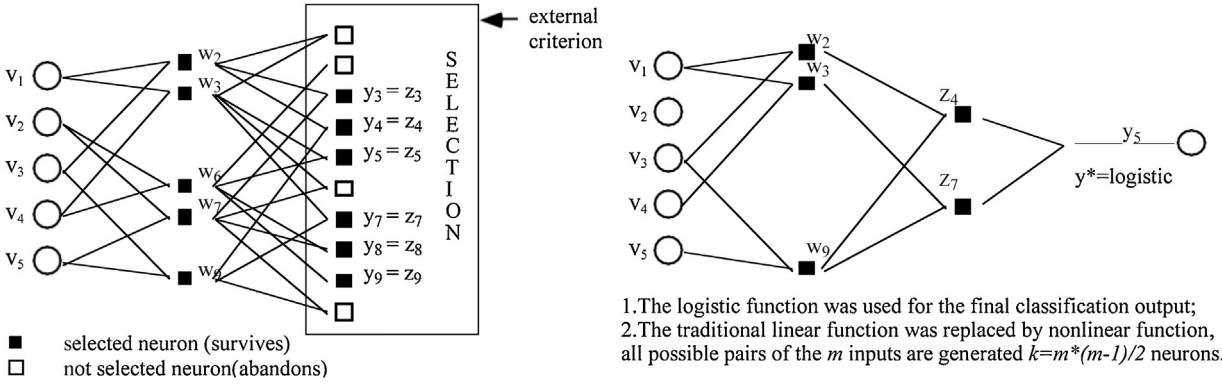

Fig. 5. The network structure of GMDH-Logistic.

Table 2
The detection result of GMDH-Logistic model.

Disease level	ID	Model detection	GMDH-Logistic prediction					
			value	class	$prob_0$	$prob_1$	$prob_2$	$prob_3$
serious disease	image01	serious disease	2.01	3	0.09	0.25	0.33	0.34
normal	image02	slight disease	1.39	1	0.22	0.29	0.27	0.23
slight disease	image03	normal	0.60	0	0.44	0.29	0.17	0.10
moderate disease	image04	moderate disease	1.60	2	0.17	0.28	0.29	0.27
normal	image05	normal	0.09	0	0.54	0.28	0.12	0.06
serious disease	image06	serious disease	2.20	3	0.06	0.23	0.33	0.37
moderate disease	Image07	serious disease	1.83	3	0.12	0.26	0.31	0.31
serious disease	image08	serious disease	2.54	3	0.03	0.20	0.34	0.42
serious disease	image09	serious disease	2.58	3	0.03	0.19	0.35	0.43
slight disease	Image10	slight disease	1.41	1	0.21	0.29	0.27	0.23
serious disease	Image11	serious disease	2.08	3	0.08	0.24	0.33	0.36
serious disease	Image12	serious disease	2.87	3	0.02	0.17	0.35	0.46
moderate disease	Image13	slight disease	1.52	1	0.19	0.28	0.28	0.25
moderate disease	Image14	serious disease	2.33	3	0.05	0.22	0.34	0.40
serious disease	Image15	serious disease	2.24	3	0.06	0.22	0.33	0.38
moderate disease	Image16	normal	0.97	0	0.34	0.29	0.21	0.15

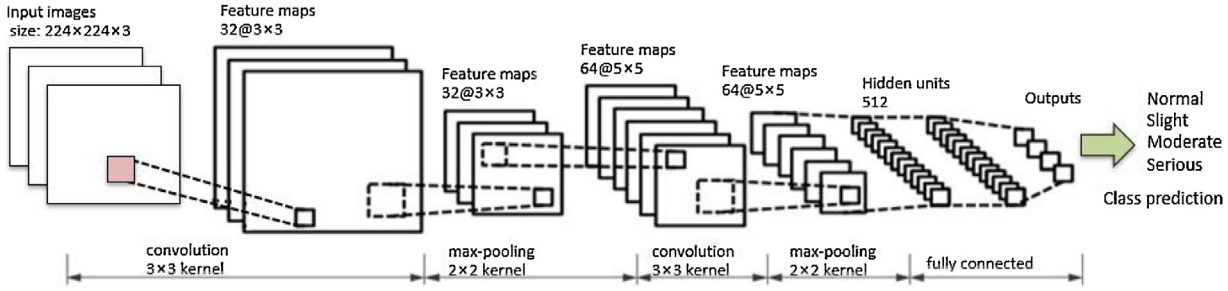

Fig. 6. The used convolutional neural network architecture.

Table 3
The predicting Comparison (unit: %).

Testing times	Algorithms					
	GMDH-Logistic	SVM	PCA-SVM	CNN	GANN	PCA-ANN
1	85.71	78.57	92.86	63.64	100.00	100.00
2	90.91	90.91	90.91	60.00	81.82	81.82
3	78.57	71.43	71.43	35.71	64.29	57.14
4	90.91	83.33	90.91	54.55	90.91	90.91
5	90.00	90.00	80.00	57.14	90.00	90.00
average	86.67	81.97	85.00	53.33	85.00	83.33

transform, the original images are augmented to at least 100 images per category for the CNN model training, and inspired by AlexNet [40], we utilized a network composed of 2 convolutional layers, 2 max-pooling layers, and 1 fully connected layer, as depicted in Fig. 6. The $224 \times 224 \times 3$ image followed by the convolutional layer with 32 kernels of size 3×3 is used as the input of the network, and each image is resized to the uniform dimension to fit the model. The results are shown as the following Table 3.

From the comparison of different algorithms, the average recall rate of the GMDH-Logistic algorithm is 86.67 %, which is greater

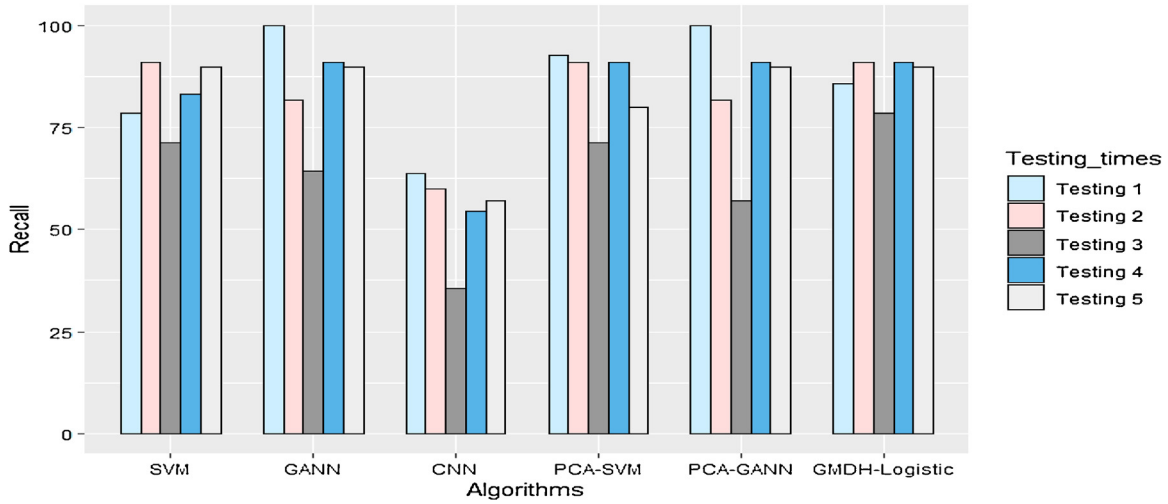

Fig. 7. The comparison of algorithm performance.

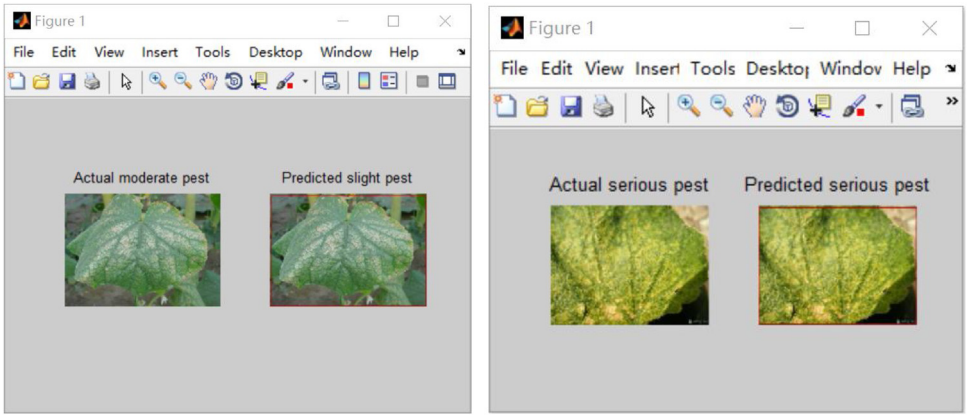

Fig. 8. The contrast images of plant disease detection.

than other algorithms. As can be seen from the above Fig. 7, except that the recall rate of the GMDH-Logistic algorithm is lower than that of other algorithms at one time, it is higher than that of other algorithms at other times. So, by contrast, this algorithm shows a better ability to distinguish whether the plant is the diseased plant or not, and the performance of the GMDH-Logistic algorithm is superior to that of other algorithms in this case. Particularly, the results were compared with CNN, which is the most popular classifier for large-scale images because of a large number of parameters that need to be trained. However, although the data augmentation techniques are applied in the CNN method, the reliable predicting results are not yielded due to the lack of a substantial number of images. By contrast, under the condition of small sample data, the proposed feature extraction based GMDH-Logistic approach shows a significant classifying effect and which is suitable for the detection of plant disease images.

On the other hand, for the specific severity categories of plant leaf disease, there may still be some departures for the proposed method, for example, the prediction probability of slight disease and moderate disease were both 0.28 for the 13th image in Table 2, and due to the negligible gap, this category was classified to the slight disease rather than the moderate disease, which causes the deviations as a result. Therefore, some small deviations can cause the difference in classifying. Additionally, there may be some ambiguity in the labeling data itself, which can cause the effect as well. Thus, this is a direction for improvement in further study. Fig. 8 depicts an example of contrast images for the actual and predicted plant disease images.

By comparing with the actual images, it can be seen that the predicted plant disease categories are basically consistent with the actual categories, which shows that the GMDH-Logistic algorithm can identify the plant diseases well, then it can assist in monitoring the growth of plants and provide a basis for producers to make decisions.

5. Conclusions

The extraction of relevant features plays an important role in the fields of image processing and computer vision, and which is the essential step of plant disease detection by leaf images. Therefore, this paper extracted the critical features of leaf images through the GIWA filtering, image segmentation, gray level co-occurrence matrix, etc. Including the *Pct*, *Num*, Contrast1, etc., 15 features were extracted in the paper, on the basis of this, the index system for the model prediction was established.

Moreover, according to the result of feature analysis, we introduced the self-organizing data mining technology in the image recognition field, and proposed a novel GMDH-Logistic method for the automatic detection and classification of plant leaf diseases. The critical features are automatically selected to enter the model, and the determined variables are usually interpretable, which overcomes the shortcomings of other algorithms.

The relevant comparative experiments are performed, and under complex background conditions, although the state-of-the-art methods such as CNN has the ability of automatic feature extraction, the proposed feature engineering based GMDH-Logistic approach presents a significant classifying effect and which is suitable for the classification of plant disease images.

Author statement

I have made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND

I have drafted the work or revised it critically for important intellectual content; AND

I have approved the final version to be published; AND

I agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

All persons who have made substantial contributions to the work reported in the manuscript, including those who provided editing and writing assistance but who are not authors, are named in the Acknowledgments section of the manuscript and have given their written permission to be named. If the manuscript does not include Acknowledgments, it is because the authors have not received substantial contributions from nonauthors.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

This work is partly supported by the grants from the National Natural Science Foundation of China (Project no. 61672439) and the Fundamental Research Funds for the Central Universities(#20720181004). The authors wish to thank all the editors and anonymous reviewers for their constructive advice.

References

[1] B. Faithpraise, C. Chatwin, Automatic plant pest detection & recognition using k- means clustering algorithm & correspondence filters, Int. J. Adv. Biotechnol. Res. 4 (2) (2013) 189–199.
[2] H. Al Hiary, S. Bani Ahmad, M. Reyalat, Z. M.Braik, A.L. Rahamneh, Fast and accurate detection and classification of plant diseases, Int. J. Comput. Appl. 17 (1) (2011) 31–38.
[3] Weiguang Ding, Graham Taylor, Automatic moth detection from trap images for pest management, Comput. Electron. Agric. 123 (2016) 17–28.
[4] Y.-Q. Hu, L.-T. Song, J. Zhang, C.-J. Xie, R. Li, Pest image recognition of multi-feature fusion based on sparse representation, Pattern Recognition and Artificial Intelligence 27 (11) (2014) 985–992.
[5] M.A. Ebrahimi, M.H. Khoshtaghaza, S. Minaei, B. Jamshidi, Vision-based pest detection based on SVM classification method, Comput. Electron. Agric. 137 (2017) 52–58.
[6] A.K. Sahu, G. Swain, Dual stego-imaging based reversible data hiding using improved LSB matching, Int. J. Intell. Eng. Syst. 12 (5) (2019) 63–73.
[7] Anand H. Kulkarni, R.K. Ashwin Patil, Applying image processing technique to detect plant diseases, Int. J. Mod. Eng. Res. 2 (5) (2012) 3661–3664.
[8] M.A. Ebrahimi, M.H. Khoshtaghaza, S. Minaei, B. Jamshidi, Vision-based pest detection based on SVM classification method, Comput. Electron. Agric. 137 (2017) 52–58.
[9] J. Garcia, C. Pope, F. Altimiras, A distributed K-means segmentation algorithm applied to lobesia botrana recognition, Complexity 14 (2017).
[10] C. Chen, Y.Q. Shi, Jpeg image steganalysis utilizing both intrablock and interblock correlations, in: IEEE International Symposium on Circuits and Systems, 2008, ISCAS 2008, IEEE, 2008, pp. 3029–3032.
[11] T. Pevny, P. Bas, J. Fridrich, Steganalysis by subtractive pixel adjacency matrix, IEEE Trans. Inf. Forensics Secur. 5 (2) (2010) 215–224.
[12] R.D.L. Pires, D.N. Gonçalves, J.P.M. Oruê, W.E.S. Kanashiro, J.F. Rodrigues Jr., B.B. Machado, W.N. Gonçalves, Local descriptors for soybean disease recognition, Comput. Electron. Agric. 125 (2016) 48–55.
[13] C. Chen, Y.Q. Shi, Jpeg image steganalysis utilizing both intrablock and interblock correlations, in: IEEE International Symposium on Circuits and Systems, ISCAS 2008, IEEE, 2008, pp. 3029–3032.
[14] M. Dash, H. Liu, Feature selection for classification, Adv. Intell. Data Anal. 1 (3) (1997) 131–156.
[15] N. Mangathayaru, B.M. Bai, P. Srikanth, Clustering and classification of effective diabetes diagnosis: computational intelligence techniques using PCA with kNN, International Conference on Information and Communication Technology for Intelligent Systems (2018) 426–440.
[16] J.V. Ral, M.N. Rao, D. Devarpalli, P. Srikanth, Identification of AIDS disease severity based on computational intelligence techniques using clonal selection algorithm, Int. J. Converg. Comput. 2 (3/4) (2016) 193–207.
[17] N. Guettari, A.S. Capelle-Laizé, P. Carré, Blind image steganalysis based on evidential k-nearest neighbors, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 2742–2746.
[18] S. Deepa, Steganalysis on images using svm with selected hybrid features of gini index feature selection algorithm, Int. J. Adv. Res. Comput. Sci. 8 (5) (2017).
[19] M. Ramezani, S. Ghaemmaghami, Towards genetic feature selection in image steganalysis, in: Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE, IEEE, 2010, pp. 1–4.
[20] M. Sheikhan, M. Pezhmanpour, M.S. Moin, Improved contourlet-based steganalysis using binary particle swarm optimization and radial basis neural networks, Neural Comput. Appl. 21 (7) (2012) 1717–1728.
[21] J. Kodovsky, J. Fridrich, V. Holub, Ensemble classifiers for steganalysis of digital media, IEEE Trans. Inf. Forensics Secur. 7 (2) (2012) 432–444.
[22] S. Phadikar, J. Sil, A. Das, Classification of rice leaf diseases based on morphological changes, Int. J. Inf. Electron. Eng. 2 (2012) 460–463.
[23] S. Kumar, et al., Plant leaf disease identification using exponential spider monkey optimization, Sustain. Comput. Inform. Syst (2018), http://dx.doi.org/10.1016/j.suscom.2018.10.004.
[24] M.A.A. Kahar, S. Mutalib, S. Abdul-Rahman, Early Detection and Classification of Paddy Diseases with Neural Networks and Fuzzy Logic, Recent Advances in Mathematical and Computational Method, 2015, pp. 248–257.
[25] A. Kamilaris, F.X. Prenafeta-Boldú, Deep learning in agriculture: a survey, Comput. Electron. Agric. 147 (2018) 70–90.
[26] N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov, Deep learning classification of land cover and crop types using remote sensing data, IEEE Geosci. Remote Sens. Lett. 14 (5) (2017) 778–782.
[27] H. Gensheng, Y. Xiaowei, Z. Yan, W. Mingzhu, Identification of tea leaf diseases by using an improved deep convolutional neural network, Sustain. Comput. Inform. Syst. 24 (2018), 100353.
[28] A. Kamilaris, F.X. Prenafeta-Boldú, Deep learning in agriculture: a survey, Comput. Electron. Agric. 147 (2018) 70–90.
[29] S.P. Mohanty, D.P. Hughes, M. Salathé, Using deep learning for image-based plant disease detection, Front. Plant Sci. 7 (2016) 1–10.
[30] R. Kawasaki, H. Uga, S. Kagiwada, H. Iyatomi, Basic study of automated diagnosis of viral plant diseases using convolutional neural networks, in: Proceedings of the International Symposium on Visual Computing (ISVC), Las Vegas, NV, USA, 2015, pp. 638–645.
[31] Jayme G.A. Barbedo, Factors influencing the use of deep learning for plant disease recognition, Biosyst. Eng. 172 (2018) 84–91.
[32] S.P. Mohanty, D.P. Hughes, M. Salathe, Using deep learning for image-based plant disease detection, Front. Plant Sci. 7 (2016), Article 1419.
[33] E. Navon, O. Miller, A. Averbuch, Color image segmentation based on adaptive local thresholds, Image Vis. Comput. 23 (1) (2005) 69–85.
[34] Li Guo, Long Chen, C.L. Philip Chen, Jin Zhou, Integrating guided filter into fuzzy clustering for noisy image segmentation, Digit. Signal Process. 83 (2018) 235–248.
[35] David Hong, Laura Balzano, Jeffrey A. Fessler, Asymptotic performance of PCA for high-dimensional heteroscedastic data, J. Multivar. Anal. 167 (2018) 435–452.
[36] Yang Liu, Shuangshuang Zhao, Qianqian Wang, Quanxue Gao, Learning more distinctive representation by enhanced PCA network, Neurocomputing 275 (2018) 924–931.
[37] Mansour Sheikhan, Najmeh Mohammadi, Neural-based electricity load forecasting using hybrid of GA and ACO for feature selection, Neural Comput. Appl. 21 (2012) 1961–1970.
[38] Mansour Sheikhan, Najmeh Mohammadi, Time series prediction using PSO-optimized neural network and hybrid feature selection algorithm for IEEE load data, Neural Comput. Appl. 23 (2013) 1185–1194.
[39] C.Z. He, J. Wu, J.A. Müller, Optimal cooperation between external criterion and data division in GMDH, Int. J. Syst. Sci. 39 (6) (2008) 601–606.
[40] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012).