Automatic difference vegetation index generator for spider mite-infested cotton detection using hyperspectral reflectance

Huasheng Huang[1,2], Jizhong Deng[3,2], Yubin Lan[4,2*], Aqing Yang[1*], Yan Jiang[5], Gaoyu Suo[3,2], Pengchao Chen[4,2]

(1. *College of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510665, China*;
2. *National Center for International Collaboration Research on Precision Agricultural Aviation Pesticide Spraying Technology, Guangzhou 510642, China*;
3. *College of Engineering, South China Agricultural University, Guangzhou 510642, China*;
4. *College of Electronic Engineering, South China Agricultural University, Guangzhou 510642, China*;
5. *Xinjiang Jiangtian Aerial Science and Technology Co., Ltd, Shihezi 832000, China*)

Abstract: Spider mites are one of the main pest stresses on cotton, which cause serious economic losses in cotton production in Xinjiang Region. This article explored the potential of ground based hyperspectral reflectance for mite-infestation detection. Also, the possibility to reduce bands and simplify analyzing was studied. In this regard, an automatic difference vegetable index was researched, which required only two bands and a simple subtraction operation. A multi-objective genetic algorithm was proposed for band selection, and its performance was compared with the mainstream machine learning methods. Experimental results showed that the proposed approach outperformed others in accuracy with less complexity. All the results revealed that the proposed method has potential in mite- infestation detection in agricultural applications.

Keywords: hyperspectral, spider mite, genetic algorithm, band selection
DOI: 10.33440/j.ijpaa.20200302.88

Citation: Huang H S, Deng J Z, Lan Y B, Yang A Q, Jiang Y, Suo G Y, Chen P C. Automatic difference vegetation index generator for spider mite- infested cotton detection using hyperspectral reflectance. Int J Precis Agric Aviat, 2020; 3(2): 83–88.

1 Introduction

Spider mites are one of the main pest stresses on cotton, which causes serious cotton yield reduction in xinjiang province in China. They feed on plants, causing leaf puckering and reddish discoloration in early stages of infestation and leaf drop later[1]. It is necessary to obtain the mite- infestation condition by remote sensing means since it is more efficient than manual field scouting[2]. However, the spider mites often locate in the middle or lower layers of cotton, which is not visible from canopy. Especially in the early stage of mite- infestation, no obvious symptoms will be revealed from the upper layer. Satellite remote sensing or aerial remote sensing cannot well recognize the mite-infestation at its early stages, since these remote sensing means can only obtain the canopy information of cotton. In contrast, ground based hyperspectral remote sensing offers the possibility for mite-recognition[3]. The potential of the ground based hyperspectral remote sensing is considered for two main reasons: (1) the hyperspectral sensors are close to the crop, which may obtain the reflectance difference caused by the pest stress; (2) high spectral resolution offers higher possibility to capture the reflectance difference at some certain bands.

In recently years, several researches are conducted to evaluate the potential of ground based hyperspectral remote sensing for mite- infestation detection. Reisig et al.[4] collected the hyperspectral reflectance of cotton using a portable hyperspectral spectrometer. Vegetable indices were calculated in order to distinguish the pest infested cotton from the healthy ones. Experimental results showed that it is possible to recognize the pest infestation by tracking the spectral changes in leaves. Herrmann et al.[5] collected the hyperspectral reflectance data of greenhouse pepper (Capsicum annuum) leaves, and transformed the reflectance into vegetation indices for classification. Experimental results showed that early identification of TSSM greenhouse pepper leaf damage can be obtained by ground based hyperspectral means.

However, the hyperspectral devices are expensive, which is impossible to be widely applied in agricultural applications. One of the solutions is to use feature selection methods to reduce the bands required, and use machine learning methods for classification. Zhang et al.[6] used the hyperspectral technique to classify the barnyard grass, weedy rice and normal rice. Hyperspectral images of 281 leaves were captured, with the spectral range from 380 to 1080 nm. Feature selection was explored through successive projection algorithm (SPA), and

Received date: 2020-04-24 **Accepted date:** 2020-06-15
Biographies: Huasheng Huang, PhD, Lecturer, research interests: UAV agricultural remote sensing, Email: huanghsheng@126.com; **Jizhong Deng,** PhD, Professor, research interests: UAV agricultural remote sensing, Email: jz-deng@scau.edu.cn; **Yan Jiang,** research interests: Precision agricultural aviation application, Email: 2803620@qq.com; **Gaoyu Suo,** Postgraduate student, research interests: Precision agricultural aviation application, Email: 447886873@qq.com; **Pengchao Chen,** Postgraduate student, research interests: Precision agricultural aviation application, Email: chenpengchao808@163.com.
***Corresponding Author: Aqing Yang**, PhD, Lecturer, research interests: Artificial Intelligence and Image Processing. Mailing Address: Department of Computer Sciences, Guangdong Polytechnic Normal University. Email: 512289446@qq.com. **Yubin Lan**, PhD, Distinguished Professor, Director, research interests: precision agricultural aviation application. Mailing Address: College of Electronic Engineering, South China Agricultural University. Email: ylan@scau.edu.cn.

classification was conducted using support vector machine (SVM) and random forests algorithms. Experimental results showed that the weighted support vector machine with 6 spectral features selected by SPA can achieve 100%, 100%, and 92% recognition rates for barnyard grass, weedy rice and rice, respectively. Liu et al.[7] collected the hyperspectral images of soybeans for variety classification. The fuzzy rough set (FRS) theory was adopted for hyperspectral band selection. The Extreme Learning Machine and Random Forest were used for classification. Experimental results showed that a subset containing eight bands achieved an average accuracy of 99.11%.

From the aforementioned literatures, it can be seen that the feasibility of using hyperspectral technique for mite- infestation recognition. However, there are some shortages for agricultural applications: (1) Most of the experiments are conducted in the greenhouse or inside the lab, which may be different from the field conditions; (2) the selected bands after band reduction is still slightly large, which will bring high costs when customizing spectral sensors; (3) most of the analyzing methods evolved are machine learning models, which are hard to be widely used by ordinary farmers. This research is conducted to overcome all these problems. The objective of this paper is to explore a solution which is easy to use, with few bands required and satisfied accuracy.

2. Materials and methods

2.1 Data collection and preprocessing

2.1.1 Data collection

The experimental sites were located in Shihezi city, Xinjiang province, China. Two cotton fields were selected for data collection, where one is infested with spider mites (44°23′44″N, 85°55′2″E) and the other is normal (44°23′45″N, 85°55′0″E), as shown in Figure 1a. Data collection was conducted on 2017 19th July, when the cotton was in its early growth stages and the mite infestation started to evolve. A field spectrometer (FieldSpec3, ASD, USA) was used to collect the spectral of the cotton leaves, as shown in Figure 1b. The spectral range of the spectrometer is 350-2500 nm with 1nm bandwidth. During data collection, the probe was consistently kept 5 cm above the leaves. The healthy condition (normal or mite- infested) for each sample was recorded during data collection, as shown in Figure 2.

a. Experimental field b. Scene of data collection

Figure 1 An illustration of the experimental field and the scene of data collection

a. Normal b. Mite-infested c. Mite-infested

Figure 2 The leaves of the normal cotton and the mite-infested ones

In the mite- infested cotton field, 110 mite- infested samples and 50 normal samples were collected. In the normal field, 27 normal samples were collected. Before data collection, the spectral of the white board was captured. After that, the spectral of each sample was collected 5 times to reduce the random error, and the averaged spectral was computed and used as the digital number of the sample. The collected spectral was demonstrated in Figure 3. From Figure 3 It can be seen that the spectra are seriously overlapping in each wavelength regions, and it can hardly find a single band to distinguish these two types. One possible reason is that mite- infestation was in its early stage, which had little influence on the biological condition of the leaves, thus the infested cotton reflected similar spectral with the normal ones.

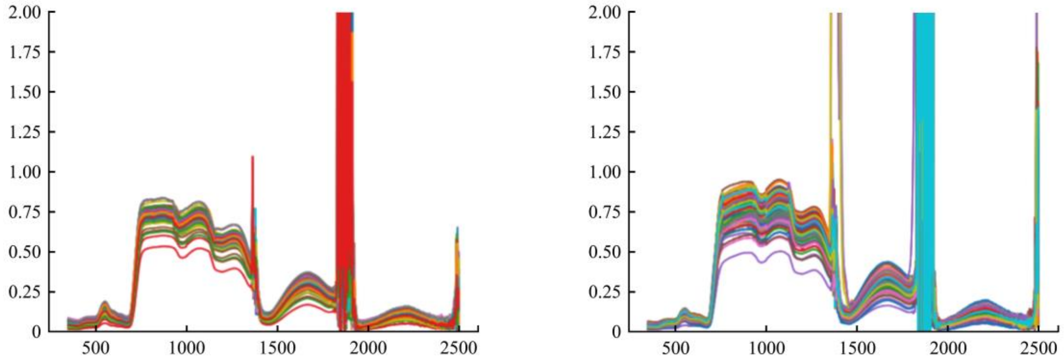
2.1.2 Data preprocessing

White calibration is necessary to avoid the illumination changes, and all spectra were converted from digital number to reflectance using the formula below:

$$S_r = \frac{S_{DN}}{S_W} \tag{1}$$

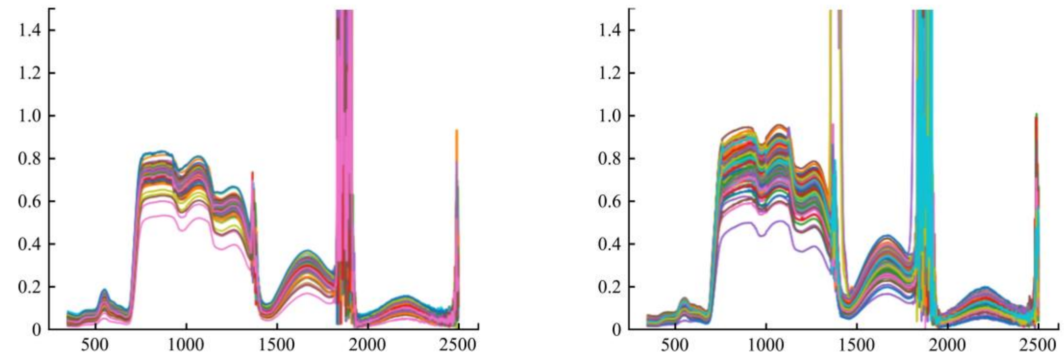
where, S_{DN} and S_r represent the digital number and the reflectance, and S_W denotes the white reflectance spectral collected before the data collection. Due to the high noise of the spectral reflectance in the spectral range of 1000-1100, 1400-1620 nm and 2000-2500 nm (Figure 4), these bands were removed and the remaining 1678 bands were considered for further analysis, as shown in Figure 5.

The dataset in this research was consisted of 187 samples, including 110 mite-infested samples and 77 normal samples. In the dataset, 50 mite-infested samples and 37 normal samples were randomly selected as training set, and the remaining 60 mite-infested samples and 40 normal samples were used as testing set.



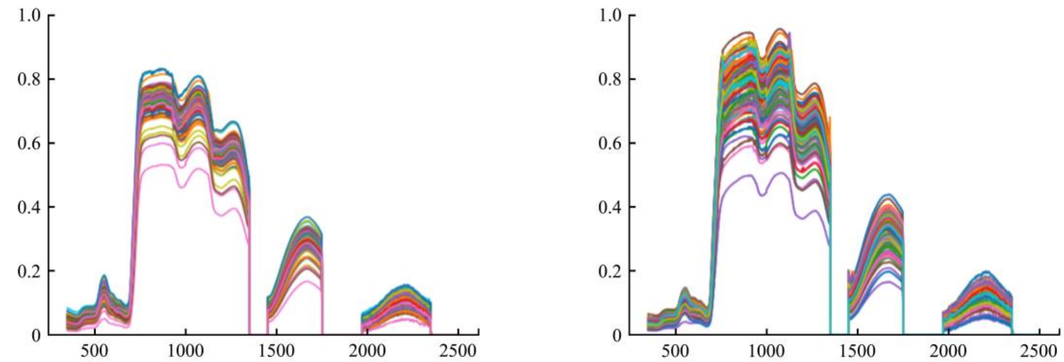
a. Spectral of normal cotton b. Spectral of mite- infested cotton

Figure 3 The collected digital number of the normal cotton and the mite- infested ones



a. Reflectance of normal cotton b. Reflectance of mite- infested cotton

Figure 4 The reflectance of the normal cotton and the mite- infested ones



a. Reflectance of normal cotton after dennise b. Reflectance of mite- infested cotton after denoise

Figure 5 Reflectance of the normal cotton and the mite- infested ones after denoise

2.2 Methods

The objective of this research was to find a smallest subset of bands with excellent accuracy. Also, the computation of the selected bands should be as simple as possible to promote its application. It can be seen from Figure 3 that the spectra of the normal cotton and infested ones are seriously overlapping in each bands, thus it is infeasible to classify using one single band. With this knowledge given, a hypothesis arises: *Is there exist a vegetable index that is capable of detecting mite- infestation with only two bands?* The hypothesis was formulated using the formula below:

$$DVI = band1 - band2 \quad (2)$$

where, *band*1 and *band*2 represent two selected bands, and *DVI* denotes their difference. If the hypothesis was substantiated, there would be two tremendous advantages: (1) the bands required are fewer than most reported literatures, which may significantly reduce the costs when customizing the corresponding spectral sensors; (2) the data implementation is easy, which may enhance its usage in agricultural applications like the widely used difference vegetable index[8]. Thus, the objective of this research can be transformed to find the two bands that support the hypothesis.

However, there are $1678\times(1678-1)\div2=1407003$ choices for the selection of two bands. It is impossible to test the performance of these selections one by one. In view of the research objective, the selection of two bands could be regarded as an optimization problem where the goal is to maximize the accuracy. Figure 6 illustrates the classification accuracy with some tested bands. It can be seen from Figure 6 that the selection of two bands is not a pure convex problem, and the gradient base approaches would be easily trapped by the local optimization points. In this case, the genetic algorithm (GA)[9] was adopted to seek for the optimal bands. It was supposed that with the randomness brought by the crossover and mutation of genetic algorithm, the problems of the gradient based methods can be well addressed.

2.2.1 Genetic algorithm

The GA randomly initializes several individuals as the original population, and employs three operators (selection, crossover, and mutation) to propagate its population from one generation to another[10]. The selection operator selects several individuals that fit the evaluation metrics best. The crossover operator produces two new individuals by recombining the information from two parents. Mutation is a random alteration of some gene values in an individual. The mutation operator allows for global search of the design space and prevents the algorithm from getting trapped in local optimization.

The GA optimization includes several generations. In each generation, the crossover and mutation operator are applied to improve the original population. After that, the optimized individuals and the original population are merged and evaluated, where the best individuals will be selected and used as the original population in the next generation. The iteration will terminate when the required accuracy or the maximum times is reached. The general workflow of GA is illustrated in Figure 7.

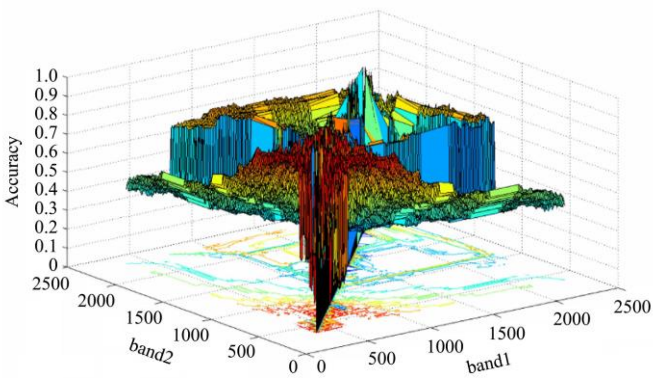


Figure 6 Classification accuracy with different bands selected

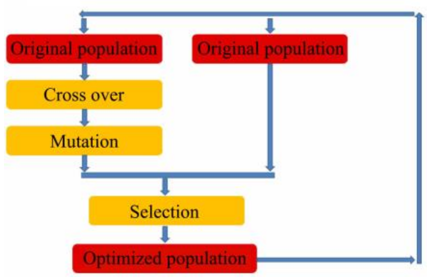


Figure 7 The workflow of genetic algorithm

2.2.2 Evaluation metrics

Like most of other literatures, we applied the classification accuracy as the metrics, which can be formulated by the formula below:

$$accuracy = \frac{num_{acc}}{num_{all}} \tag{3}$$

where, num_{acc} represents the number of samples been correctly classified, and num_{all} denotes the number of all samples. The modeling process is to find a separation line for the DVI value (formula 1). The generated vegetable index could achieve 100 percent accuracy when the separation line that can completely separate the DVI values of different types, as shown in Figure 8.

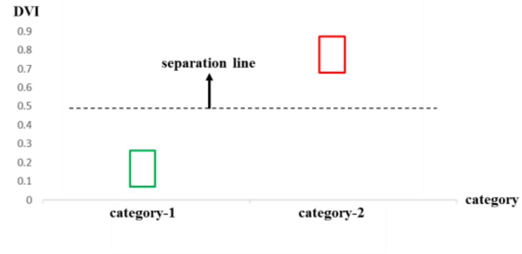


Figure 8 The separation line of DVI

Besides the accuracy, we also applied the margin as the metrics. The separation margin refers to the difference between the upper border of the first category and the lower boarder of the second category, as shown in Figure 9. The idea of this metrics came from the optimization goal of support vector machines (SVM) [11], where a larger separation margin indicates a smaller structure risk. Besides that, the mean value of the upper border of the first category and the lower boarder of the second category would be considered as the separation threshold.

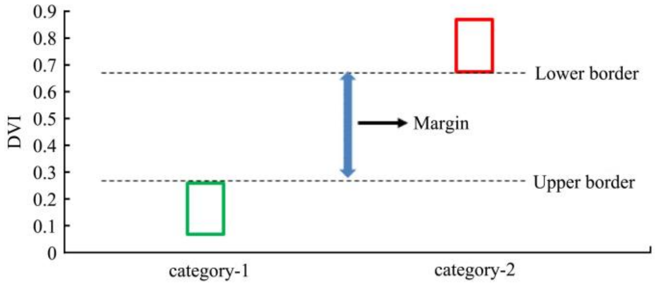


Figure 9 The margin of the DVI

During the evaluation in each generation of the GA, the accuracy was considered as the first metrics. Under the circumstances with the same accuracy, the separation margin would be considered as the second metrics, in order to achieve better generation capability. In this case, the method we proposed is a multi-objective genetic algorithm.

3 Results and discussion

3.1 Experiments on GA

In this section, the GA method was used to seek for two optimal bands, and their difference was used for classification. There are many versions to improve the GA method. In this research, a basic binary encoded GA with tournament selection, uniform crossover and low probability mutation rate is employed. Therefore, the original population was initialized in binary format, while the optimized results were decoded into decimals. The population and the max generalization were both set to 100. There is no setting for the optimal fitness value, in order to get a better result. The probability for the crossover and mutation were set to 0.8 and 0.01. The maximum generation was set to 200, which is found sufficient for the optimization. The GA method was run on the training set to find two optimal bands. During the optimization process, the changes on the accuracy and margin was illustrated in Figure 10. It can be seen from Figure 10 that the applied model quickly converges with excellent accuracy and margin. The selected bands were 1131 and 769, and the separation threshold was 0.0191. We applied these setting on the testing set, where the obtained accuracy and margin were 100%

and 0.0198. This experimental results showed that the proposed method can automatically generate a difference vegetable index to distinguish the normal cotton from the mite- infested ones.

Since the initialization, crossover and mutation process were performed with randomness, different GA experiment may yield different results. In order to explore the stability of this method, we conducted 10 consecutive GA experiments. The experimental results were showed in Table 1, where the accuracy and margin were evaluated on the testing set. It can be seen from Table 1 that the performance of the GA method is consistent through different experiments. Also, the selected bands in Table 1 locates in the same spectral ranges, and the experimental results were all proximate to the optimal value. Besides that, the experimental results were also explainable. The first band (1120-1131nm) was located in the NIR range, and the second band (758-773 nm) was distributed within the Red range. The mite- infestation will result in an impact on the difference of the NIR and Red bands, which is similar the principle of the NDVI[12].

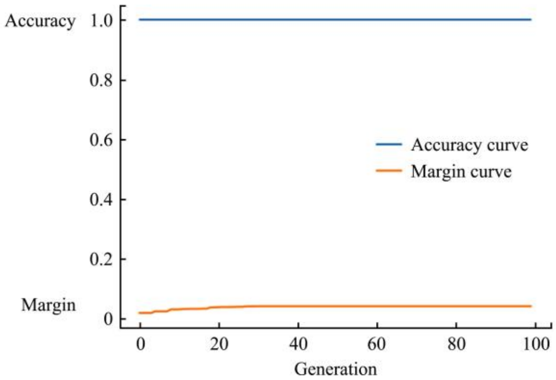


Figure 10 The accuracy and margin curves during the optimization process

Table 1 Results across different GA experiments

Experiment number	Band1	Band2	Accuracy/%	Margin
1	1131	769	100	0.0198
2	1129	759	100	0.0247
3	1129	769	100	0.0206
4	1123	773	100	0.0225
5	1122	758	100	0.0276
6	1123	758	100	0.0276
7	1123	773	100	0.0225
8	1120	758	100	0.0269
9	1120	758	100	0.0269
10	1123	757	100	0.0275

3.2 Experiments on mainstream methods

According to most literatures[13-15], the mainstream methods for hyperspectral classification problems would be the machine learning models combined with the feature selection method. In this section, we applied the popular SVM[16] and random forests[17] algorithms for classification, and used the GA method for feature selection. For the SVM method, both linear and Radial Basic Function (RBF) were used as the kernel functions. For each kernel function of the SVM model, C was set as a list of [1, 10, 100, 1000], and gamma was set as a list of [0.001, 0.0001] only in case of RBF as kernel type. For the random forests, the bootstrap samples are used. The number of tress are set between 10 to 100 with an increment of 10, and the number of selected features to build trees is the root of the spectral numbers (1678). Experimental results on the testing set are shown in Table 2. It can be seen from Table 2 that the machine learning algorithms can achieve good accuracy, slightly lower than the DVI generated by the GA method.

Table 2 Results of machine learning methods using all bands

Models	Accuracy/%
SVM-linear	99
SVM-RBF	99
Random forests	98
ours	100

In order to reduce the number of features, we applied the GA method with the machine learning algorithms. The GA method was also applied to seek for two optimal bands, which is used as features for classifiers. The hyperparameter settings and selection are same as the previous experiment, and the results are shown in Table 3. With limited input features, the linear based SVM cannot find an effective separation plane, which result in low accuracy (60%). The RBF mapped the input features into high dimensional feature space, which is more feasible to be classified with a super plane. The random forests used several decision trees to vote for the final results, which further reduce the bias with the accuracy of 98%. Though the random forests achieve the approximate accuracy with the DVI generated by GA, it is more complicated and hard to be applied by ordinary users.

Table 3 Results of machine learning methods combined with GA

Models	Band1	Band2	Accuracy/%
SVM-linear	1173	580	60
SVM-RBF	780	776	81
Random forests	1241	188	98
ours	1131	769	100

From the experimental results above, it can be concluded that the method proposed in this paper can effectively distinguish the spider mite- infested cotton using only two bands and a simple subtraction operation. Also, the proposed method outperformed the classical SVM and random forests in accuracy. Thus, the hypothesis on the difference vegetable index with only two bands was substantiated.

4 Conclusions

In this research, the hyperspectral reflectance of the spider mite-infested and normal cotton were collected in two cotton fields. During data collection, the health condition of each sample was recorded. From the application aspect, we proposed a hypothesis that there may exist a difference vegetable index that can distinguish the mite infestation with only two bands. A multi-objective GA method was proposed to seek for two optimal bands. Experimental results illustrated the generated difference vegetable index is effective in mite- infestation classification with high accuracy. We also applied the mainstream machine learning algorithms on our dataset. Comparison results demonstrated that our proposed method outperformed the mainstream machine learning approaches in accuracy, substantiating the effectiveness of our hypothesis.

There are two main advantages of the experimental results. The first is that the bands required is few, which could reduce the costs when customizing specific spectral sensors for spider mite detection. The second is that the computation is easy that would promote its application in agricultural management. However, more data in different cotton fields in different growth stages is needed to further promote and evaluate our method. Besides the mite- infestation classification, a further study should explore the potential of recognizing the severity degrees caused by mite-infestation. All these issues would be left as the future work of our study.

Acknowledgements

This work was supported by the Science and Technology Planning Project of Guangdong Province, China (Grant No.2017A020208046), the National Key Research and Development Plan, China (Grant No. 2016YFD0200700), the National Natural Science Fund, China (Grant No. 61675003), the Science and Technology Planning Project of Guangdong Province, China (Grant No. 2017B010117010), and the Science and Technology Planning Project of Guangzhou city, China (Grant No.201707010047).

[References]

[1] Szendrei Z, Strand L L, Rude P A. Integrated Pest Management for Potatoes in the Western United States. 2nd Edition, 1986; 66(8): 1784–1796. doi: 10.1016/j.socscimed.2008.01.001.

[2] Huang H, Deng J, Lan Y, Yang A, Deng X, Zhang L, et al. A two-stage classification approach for the detection of spider mite- infested cotton using UAV multispectral imagery. *Remote Sensing Letters*, 2018; 9(10): 933–941. doi: 10.1080/2150704X.2018.1498600.

[3] Martin D E, Latheef M A. Remote sensing evaluation of two-spotted spider mite damage on greenhouse cotton. *Journal of Visualized Experiments Jove*, 2017; (122). doi: 10.3791/54314.

[4] Reisig D, Godfrey L. Spectral Response of Cotton Aphid– (Homoptera: Aphididae) and Spider Mite– (Acari: Tetranychidae) Infested Cotton: Controlled Studies. *Environmental Entomology*, 2007; 36(6): 1466–1474. doi: 10.1603/0046-225X(2007)36.

[5] Herrmann I, Berenstein M, Sade A, Karnieli A, Bonfil D J, Weintraub P G. Spectral monitoring of two-spotted spider mite damage to pepper leaves. *Remote Sensing Letters*, 2012; 3(4): 277–283. doi: 10.1080/ 01431161.2011.576709.

[6] Zhang Y, Gao J, Cen H, Lu Y, Yu X, He Y, Pieters J G. Automated spectral feature extraction from hyperspectral images to differentiate weedy rice and barnyard grass from a rice crop. *Computers and Electronics in Agriculture*, 2019; 159, 42–49. doi: 10.1016/j.compag.2019.02.018.

[7] Liu Y, Wu T, Yang J, Tan K, Wang S. Hyperspectral band selection for soybean classification based on information measure in FRS theory. *Biosystems Engineering*, 2019. doi: 10.1016/j.biosystemseng.2018.12.002.

[8] Shibayama M, Akiyama T. Seasonal visible, near-infrared and mid-infrared spectra of rice canopies in relation to LAI and above-ground dry phytomass. *Remote Sensing of Environment*, 1989; 27(2): 119–127. doi: 10.1016/0034-4257(89)90011-4.

[9] Juang C F. A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*, 2004; 34(2): 997–1006. doi: 10.1109/tsmcb.2003.818557.

[10] Hassan R, Cohanim B, Weck O D. In *A Comparison of Particle Swarm Optimization and the Genetic Algorithm*, Proceedings of the 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Material Conference, 2005; 2005. doi: 10.2514/6.2005-1897.

[11] V., D. S. A., Advanced support vector machines and kernel methods. *Neurocomputing*, 2003; 55(1-2): 5–20. doi: 10.1016/s0925-2312(03)00373-4.

[12] Defries R S, Townshend J R G. NDVI-derived land cover classifications at a global scale. *International Journal of Remote Sensing*, 1994; 15(17): 3567–3586. doi: 10.1080/01431169408954345.

[13] Deng X, Huang Z, Zheng Z, Lan Y, Dai F. Field detection and classification of citrus Huanglongbing based on hyperspectral reflectance. Computers and Electronics in Agriculture, 2019; 167, 105006. doi: 10.1016/j.compag.2019.105006.

[14] Liu Z Y, Wu H, Huang J. Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis. *Computers and Electronics in Agriculture* 2010; 72(2): 99–106. doi: 10.1016/j.compag.2010.03.003.

[15] Rumpf T, Mahlein A K, Steiner U, Oerke E C, Dehne H W, Plümer L. Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 2010; 74(1): 91–99. doi: 10.1016/j.compag.2010.06.009.

[16] Chih-Chung; Chang; Chih-Jen; Lin, LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. doi: 10.1145/1961189.1961199.

[17] Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. *Isprs Journal of Photogrammetry & Remote Sensing*, 2016; 114, 24–31. doi: 10.1016/j.isprsjprs.2016.01.011.