



Feasibility of Hybrid PSO-ANN Model for Identifying Soybean Diseases

Miaomiao Ji, Northeast Agricultural University, China

 <https://orcid.org/0000-0002-1767-9166>

Peng Liu, Northeast Agricultural University, China

Qiufeng Wu, Northeast Agricultural University, China

 <https://orcid.org/0000-0002-4787-2549>

ABSTRACT

Soybean disease has become one of the vital factors restricting the sustainable development of a high-yield and high-quality soybean industry. A hybrid artificial neural network (ANN) model optimized via particle swarm optimization (PSO) algorithm, which is denoted as PSO-ANN, is proposed in this paper for soybean disease identification based on categorical feature inputs. Augmentation dataset is created via synthetic minority over-sampling technique (SMOTE) to deal with quantitative insufficiency and categorical unbalance of the dataset. PSO algorithm is used to optimize the parameters in ANN, including the activation function, the number of hidden layers, the number of neurons in each hidden layer, and the optimizer. In the end, ANN model with 2 hidden layers, 63 and 61 neurons in hidden layers respectively; Relu activation function; and Adam optimizer yields the best overall test accuracy of 92.00%, compared with traditional machine learning methods. PSO-ANN shows superiority on various evaluation metrics, which may have great potential in crop disease control for modern agriculture.

KEYWORDS

Artificial Neural Network (ANN), Machine Learning, Particle Swarm Optimization (PSO), Soybean Diseases Identification, Synthetic Minority Over-Sampling Technique

1. INTRODUCTION

Soybean, as one of the important grain and oil crop in the world, plays an important role in the world's agricultural production and trade (Wu, Zhang, & Meng, 2019). However, various soybean diseases have constrained sustainable development of high-yield and high-quality soybean industry for a long time. For one thing, there will likely be a large increase in the demand for soybean with the growth of population and economy. For another thing, soybean diseases have characteristics of large variety, great impact and local outbreaks, which have been responsible for productivity and quantitative losses in crop yield. Thus, time-saving and high-efficiency identification of soybean diseases is urgently needed.

Classification algorithms play a substantial role in crop diseases identification. With the development of sophisticated instruments and fast computational techniques, the application of machine learning technologies to diagnose crop diseases has become one of the important research

DOI: 10.4018/IJCINI.290328

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

contents of intelligent agriculture. Although advances in science and technology now make it possible for computer vision approaches to assist us in automatic detection crop diseases tasks and remarkable performances have been achieved, these classification methods based on deep learning are limited to using large amount of high-quality image data and depending on great computational capacity. Our research is devoted to using simple and efficient classification method and few diseased samples for soybean diseases identification task. Symptom features can be conveyed by plant organs such as seeds, leaves and fruits, which are generally sensitive to the state of crops and whose shape, texture and color usually contain rich information. And thus the identification of soybean diseases can be based on descriptive data. Compared with image-based methods, the statistical pathological inputs can also achieve high recognition accuracy but fast evaluation speed.

ANNs use supervised learning to determine a complex, nonlinear, multidimensional mathematical fitting and have attracted a great deal of attention recently. ANNs have so excellent generalization ability and robustness that they excel in many areas: pattern classification, function approximation, intelligent control, fault detection, signal processing and system analysis etc. In addition, ANNs have also achieved impressive results in the field of agriculture and benefit more smallholders and horticultural workers, including weeds recognition (Ahmad et al., 2018), plant diseases prediction (Sharma, Singh, & Singh, 2018), soil parameters estimation (Estrada-López, Castillo-Atoche, Vázquez-Castillo, & Sánchez-Sinencio, 2018) and land cover detection (Zhan, Tian, & Tian, 2019) etc. This paper presents an application of PSO-based multi-layer perceptron (MLP) neural network in soybean diseases identification. The proposed method is capable of providing a reliable and fast estimation of soybean diseases types on the basis of typical symptom features. The performance of the proposed technique is analyzed by comparing the classification results with the traditional machine learning methods including logistic regression (LR), k-NearestNeighbor (KNN) and support vector machine (SVM) for the same hold-out test dataset.

2. RELATED WORKS

ANNs have been widely used for plant disease recognition and prediction. Daniel et al. (2017) used ANN to predict the value of the area under the disease progress curve for the tomato late blight pathosystem. ANN demonstrated superior performance compared to conventional methods. Sharma et al. (2018) predicted the late blight disease manifestation in potato using ANN on the basis of weather parameters including maximum temperature, minimum temperature, maximum humidity, minimum humidity and rainfall. Disease severity was successfully classified and the prediction accuracy of 90.9% was achieved. P. Ahmadi et al. (2017) applied ANN analysis technique for discriminating and classifying fungal infections in oil palm trees based on spectral signatures. Their results shown that ANN was accurate in predicting the diseases and the accuracy ranged from 83.3% to 100% for different samples. However, ANNs have been suffering from a number of restrictions in the learning procedure, such as appropriate selection of parameters combination, optimal adjusting of neurons connection weights and uncertainty of global minimum convergence. The mentioned inherent drawbacks can be modified and minimized by combining the ANNs with robust optimization algorithms to deal with complicated nonlinear problems. For example, Ruiz et al. (2018) proposed an Elman neural network (ENN) for forecasting energy consumption and used a genetic algorithm (GA) to optimize the weight of the models. Their results demonstrated that GA has proven to be a useful and a key factor for optimizing ANN. Pan (2012) proposed a new fruit fly optimization algorithm (FOA) with a real application in finding maximal value and minimal value to improve general regression neural network model. From their analysis result, it could be seen that through the FOA, the spread value of the parameters in ANN was optimized, and the classification prediction capability of the ANN was obviously enhanced.

PSO also can improve and optimize a candidate solution iteratively with respect to a certain degree of quality. Many researchers have tried to improve the performance and generalization capabilities of

ANNs in scientific and engineering applications by using PSO algorithm. The combined evolutionary method, i.e. PSO-ANN, has been applied to estimate rock strength (Mohamad, Armaghani, Momeni, Yazdavar, & Ebrahimi, 2018), to predict material removal rate and surface roughness (Babu, Karthikeyan, & Punitha, 2019), to forecast solar space heating system parameters (Jamali, Rasekh, Jamadi, Gandomkar, & Makiabadi, 2019), and to model nonlinear relationship between water amount of natural gas and wide ranges of CO₂ and H₂S contents, temperature and pressure (Ahmadi, Ahmad, Phung, Kashiwao, & Bahadori, 2016). PSO-based ANN can be generated in various ways and serve various classification and regression tasks. PSO can be used for an effective training the weights and biases between layers. Upendar et al. (2010) used PSO for an effective training the connection weight matrix between layers in ANN for predicting the type of fault in electric power system. Compared with the back-propagation ANN, the proposed PSO-based multi-layer ANN was capable of producing fast and more accurate results and obtained 99.91% fault classification accuracy. Hasanipanah et al. (2016) presented a new hybrid model of ANN optimized by PSO for prediction of maximum surface settlement (MSS). They developed a predictive model for MSS prediction by incorporating PSO algorithm to minimize cost function by adjusting the weights and biases. Their results indicated that the proposed PSO-ANN model was able to predict MSS with a higher degree of accuracy in comparison with the pre-developed ANN results. PSO functionalities can be used to detect subset of features to accomplish improved classification performance than using entire features dataset (Patil & Tamane, 2020). Moreover, PSO algorithm can also be used to select optimal pentameters in ANN and our research is an example. Ebrahinzade et al. (2019) proposed a combined method of the ANN and PSO to predict the leaching efficiency of cobalt from spent lithium-ion batteries (LIBs). PSO algorithm was used to select different neuron numbers and activation functions in the training phase and root mean square error (MSE) was used as the statistical measure for determination of the optimal PSO-ANN model. The superiority of PSO-ANN models was validated by statistical thresholds compared with common ANN technique. Additionally, the satisfactory results of PSO coupled ANN in estimating black plastic types shown another evidence of its applicability in the modeling of various processes (Roh, Oh, Park, & Choi, 2017).

The performance of PSO-ANN in the classification and prediction of processes characteristics is acceptable and thus, the technique can be introduced to the comprehensive analysis and modeling of the soybean diseases identification task. In this paper, the main objective of the PSO algorithm application is to optimize parameters in ANNs to improve its performance for soybean diseases identification task.

3. MATERIALS AND METHODS

3.1 Dataset Description and Preprocessing

The dataset used for evaluating the proposed PSO-ANN method is available from UCI Machine Learning Repository (Michalski, 1980). As shown in Table 1, the soybean dataset ([http://archive.ics.uci.edu/ml/datasets/Soybean+\(Large\)](http://archive.ics.uci.edu/ml/datasets/Soybean+(Large))) used for soybean diseases types identification is composed of 35 categorical attributes, some nominal and some ordered, which describe the condition of the environment and different features of the soybean, such as the plant stand, the leaves and the seed, etc. There are 19 diseases types, i.e., diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot, diaporthe-pod-and-stem-blight, cyst-nematode, 2-4-d-injury, herbicide-injury. Only the first 15 of which have been used in our work because the last four classes have so few example samples and many missing values. Before the implementation of experiments, there are a few elements need to be preprocessed. Since there is a significant portion of missing values in the dataset, the instances cannot be simply deleted, thus the missing values are imputed with the mode of each feature rather

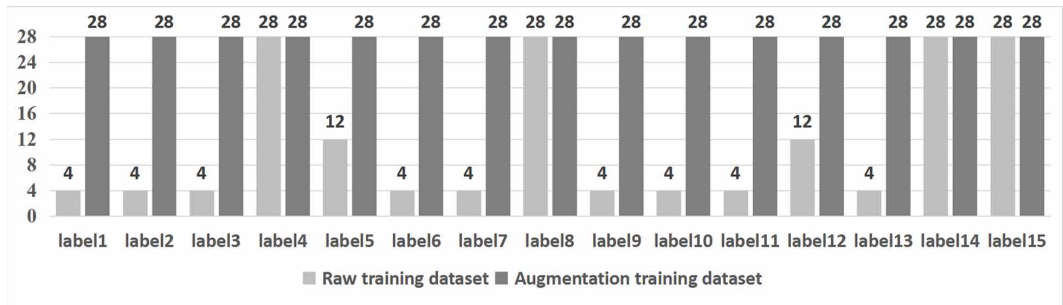
Table 1. Details of soybean dataset.

Number	Characteristic name	Contains
1	Date	April, May, June, July, August, September, October
2	Plant-Stalso	Normal, Lt-Normal
3	Precip	Lt-Norm, Norm, Gt-Standard
4	Temp	Lt-Norm, Norm, Gt-Standard
5	Hail	Yes, No
6	Crop-Hist	Diff-Lst-Year, Same-Lst-Yr, Same-Lst-Two-Yrs, Same-Lst-Sev-Yrs
7	Area-Damaged	Scattered, Low-Areas, Upper-Areas, Whole-Field
8	Severity	Minor, Pot-Severe, Severe
9	Seed-Tmt	None, Fungicide, Other
10	Germination	90-100, 80-89, Lt-80
11	Plant-Growth	Norm, Abnorm
12	Leaves	Norm, Abnorm
13	Leafspots-Halo	Absent, Yellow-Halos, No-Yellow-Halos
14	Leafspots-Marg	W-S-Marg, No-W-S-Marg, Dna
15	Leafspot-Size	Lt-1/8, Gt-1/8, Dna
16	Leaf-Shread	Absent, Present
17	Leaf-Malf	Absent, Present
18	Leaf-Mild	Absent, Upper-Surf, Lower-Surf
19	Stem	Norm, Abnorm
20	Lodging	Yes, No
21	Stem-Cankers	Absent, Below-Soil, Above-Soil, Above-Sec-Nde
22	Canker-Lesion	Absent, Below-Soil, Above-Soil, Above-Sec-Nde
23	Fruiting-Bodies	Dna, Brown, Dk-Brown-Blk, Tan
24	External-Decay	Absent, Firm-And-Dry, Watery
25	Mycelium	Absent, Present
26	Int-Discolor	None, Brown, Black
27	Sclerotia	Absent, Present
28	Fruit-Pods	Norm, Diseased, Few-Present, Dna
29	Fruit-Spots	Absent, Colored, Brown-W/Blk-Specks, Distort, Dna
30	Seed	Norm, Abnorm
31	Mold-Growth	Absent, Present
32	Seed-Discolor	Absent, Present
33	Seed-Size	Norm, Lt-Norm
34	Shriveling	Absent, Present
35	Roots	Norm, Rotted, Galls-Cysts

than the mean as the data is categorical. All the values are encoded numerically to enable the training and comparison. The dataset is shuffled in order to allow more meaningful learning .

The evaluation of the different experimental configurations is based on a train-validation-test scheme. The raw dataset is firstly divided into training data and test data. The random samples of 5 in each diseases types are selected for testing and the rest samples are used for training. To evaluate model performance and adjust model parameters, the training dataset is further split into training dataset (80%) and validation dataset (20%). Thus the raw training dataset is 172 samples in total, validation dataset is 43 samples in total and test dataset is 75 samples in total. To increase data volume and eliminate imbalances between categories, data augmentation procedure is performed on training data via SMOTE algorithm (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), which can analyze a small number of samples and artificially synthesize new samples based on the small number of samples to add to the training dataset. Depending on the amount of oversampling required, the nearest neighbors are randomly selected to produce synthetic data from the actual dataset. In this paper, the nearest three neighbors are used (Arslan, Güzel, Demirci, & Ozdemir, 2019). To the end, augmentation dataset is created, where samples in each diseases types are added to 28, respectively. And thus training dataset is 420 samples in total in augmentation dataset, which can be seen from Figure 1.

Figure 1. Data augmentation procedure via SMOTE algorithm.



3.2 ANN for Classification

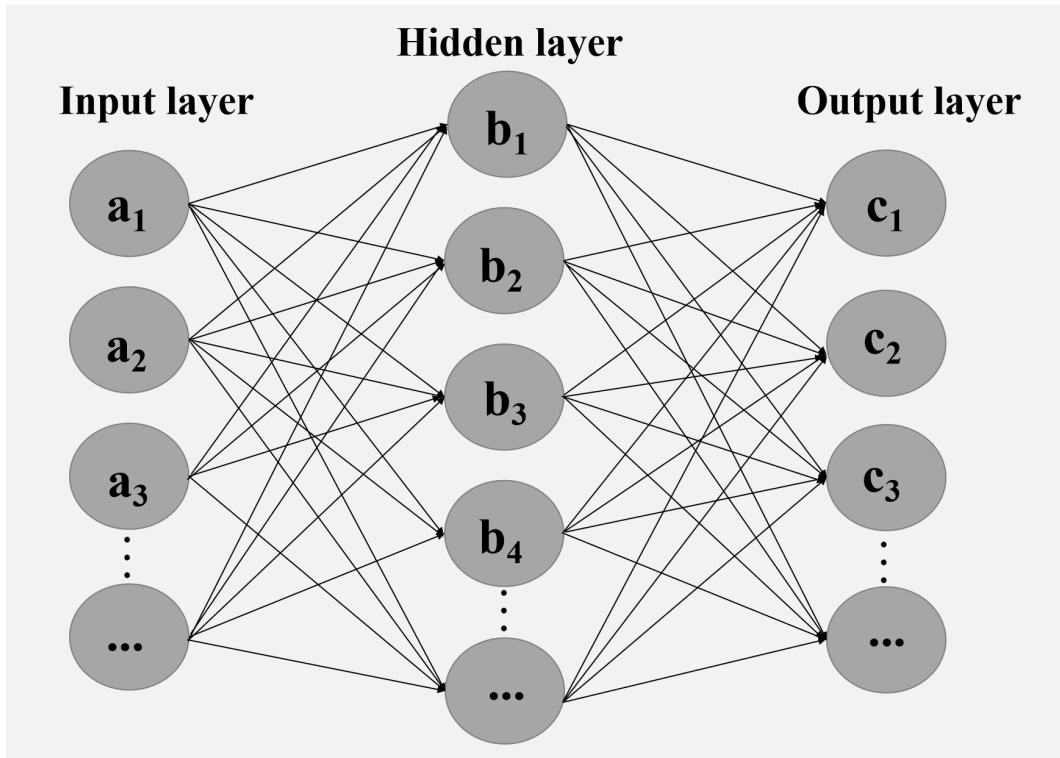
ANNs are computing units inspired by the biological complex neural structures. ANNs are notable for strong learning ability and generalization, tolerance to errors and low computational cost. MLP neural network is the most popular ANN. In the ANN structure, inputs and outputs are directly related to the input and output layers, respectively. The hidden layers are placed between the input and output layers. The transformation of information from the one layer's neuron to the other neuron of the subsequent layer conducted in the base of the following relationship (Ebrahimzade, Khayati, & Schaffie, 2019):

$$y = f\left(\sum_{i=1}^n \omega_i x_i + b\right) \quad (1)$$

where x_i is input, b is bias, ω_i is weight of neuron connection, $f(\cdot)$ is activation function, and y is output. ANN training steps are accomplished through an iterative weighting process from input to output neurons, which are called epochs. After each iteration, the evaluated output is compared with the target output based on the average accuracy. In this process, the weights and biases are modified according to back-propagation to maximize the average accuracy during the iteration. The configuration of the neural network includes the number of neurons in the input layer, the hidden

layer, the output layer, the connection weights and bias, and the activation function. The specific structure is shown in Figure 2. After the training phase ends, the validation and testing phase can be performed using unused data.

Figure 2. The standard ANN structure.



3.3 PSO for Optimization

PSO algorithm inspired by society and self-cognitive behavior was proposed by Kennedy and Eberhart (Eberhart & Kennedy, 1995). PSO is a populated search method for the optimization of continuous nonlinear functions resembling the movement of organisms in a bird flock or fish school. PSO has advantages such as easy implementation, low parameters, and high convergence rate. PSO can be effectively used in various areas, like medical data processing, machine learning, pattern matching and feature selection. As shown in Figure 3, the flow diagram of the PSO algorithm is provided below (Da & Xiurun, 2005).

Step 1: Initialization of position and velocity.

A population of random potential solution is generated in the searching space. Let $X_i = X_{i,1}, X_{i,2}, \dots, X_{i,j}$ be the position components and $V_i = V_{i,1}, V_{i,2}, \dots, V_{i,j}$ be the velocity components, where $i = 1, 2, \dots, N$, $j = 1, 2, \dots, M$, N is the the number of particles and M is the number of input variables to be optimized.

Step 2: Updating of position and velocity.

In each iteration, the new position of each particle can be worked out:

$$X_{i,j}(t+1) = X_{i,j}(t) + V_{i,j}(t+1) \quad (2)$$

where t is the current iteration number and the new velocity of each particle can be calculated according to the following formula:

$$V_{i,j}(t+1) = \omega V_{i,j}(t) + C_1 r_1 (P_{i,j}^{best} - P_{i,j}(t)) + C_2 r_2 (G_{i,j}^{best} - G_{i,j}(t)) \quad (3)$$

where $P_{i,j}^{best}$ is the best one of the solutions this particle has reached, $G_{i,j}^{best}$ is the best one of the solutions all the particles have reached. ω is the inertia weight, C_1 and C_2 are the acceleration constants, r_1 and r_2 are random numbers in the range $[0, 1]$.

Step 3: Updating of $P_{i,j}^{best}$ and $G_{i,j}^{best}$.

The fitness function $F(\cdot)$ as the objective function is used to calculate the optimum $P_{i,j}^{best}$ and $G_{i,j}^{best}$. Considering the maximization problem, the personal and global best positions at the next iteration are defined as:

$$P_{i,j}^{best}(t+1) = \begin{cases} X_{i,j}(t+1), & F(X_{i,j}(t+1)) \geq F(P_{i,j}^{best}(t)) \\ P_{i,j}^{best}(t), & F(X_{i,j}(t+1)) < F(P_{i,j}^{best}(t)) \end{cases} \quad (4)$$

$$\begin{aligned} G_{i,j}^{best}(t+1) &\in \left\{ P_{i,j}^{best}(0), P_{i,j}^{best}(1), \dots, P_{i,j}^{best}(t) \right\} \Big| F(G_{i,j}^{best}(t+1)) \\ &= \min \left\{ F(P_{i,j}^{best}(0)), F(P_{i,j}^{best}(1)), \dots, F(P_{i,j}^{best}(t)) \right\} \end{aligned} \quad (5)$$

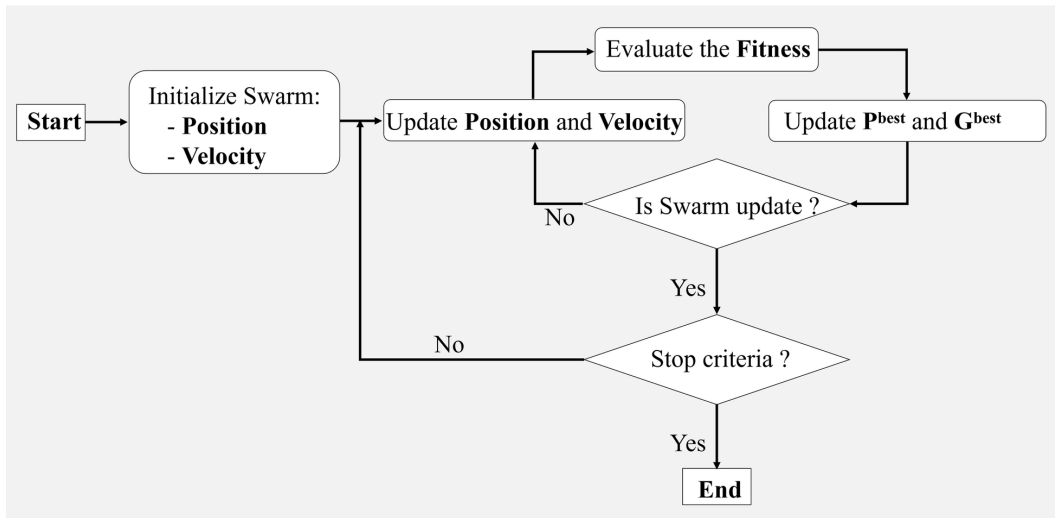
Step 4: Termination checking.

The algorithm repeats Steps 2-3 until the iteration termination condition is met and an optimal solution for the problem can be obtained. Once terminated, the algorithm reports the values of $G_{i,j}^{best}$ and $F(G_{i,j}^{best})$ as its solution.

3.4 PSO-Based ANN for Soybean Diseases Identification

The ANN technique offers outstanding potential of learning algorithm and matching of input and output association on account of its ability to recognize the relations at the back of the complex processes. The performance of the ANN models depends strongly on the parameters setting. Therefore, determination of the optimal architecture is required to design an ANN model. In this application, one major difficulty is to determine the appropriate number of hidden layers. Hidden

Figure 3. The PSO algorithm in standard flow chart.



layer is responsible for the internal representation of the data and the information transformation between input and output layers.

The hidden layer can have multiple layers to enhance the expressiveness of the model. But it also increases the model computation complexity. To achieve good performance, it is necessary to evaluate various values of hidden layers. Hence, single-layered ANN, double-layered ANN and multi-layered ANN with three hidden layers are considered, which are denoted as PSO-ANN_1, PSO-ANN_2, PSO-ANN_3, respectively. In the next stage of ANN design, the optimum design for the number of neurons in the hidden layer is required. If there are too few neurons in the hidden layer, the network may not contain sufficient degrees of freedom to form a representation. If too many neurons are used, the network might become over trained. Therefore, to achieve good performance, it is necessary to evaluate for various values of neurons in hidden layer. The number of hidden layer neurons is set to be in range specified by minimum 3 and maximum 200 in our research.

In addition, determining activation function in the hidden layer is also a critical task in the ANN architecture, which directly affects the efficiency and stability of training. The activation function can introduce non-linear factors into the neurons, so that ANN can approach any non-linear function arbitrarily. Optimizer algorithm used to update the weights and bias is another important parameter that affects the convergence speed and classification accuracy score. The choices of activation functions include Relu, Tanh and Logistic and the choices of optimizer include SGD, Adam and Lbfgs.

3.5 Experimental Setup

The proposed models is built using the open source Sklearn package on top of Python environment. To enable a fair comparison between the results, all the other parameters are default and identical across all the experiment configurations. To improve model generalization and avoid randomness, the fitness function of hybrid PSO-ANN model for soybean diseases identification is obtained according to the maximum value of average validation accuracy trained ten times. Notice that the training samples and the validation samples should be randomly shuffled based on the fixed ratio of 8: 2 each time. The values of C_1 and C_2 in PSO algorithm, the number of particles N are selected 2, 2 and 20, respectively. Each of these experiments runs for a total of 25 iterations and the training processes are recorded. The model with the best fitness value is saved and will be used for prediction and evaluation on test dataset.

3.6 Evaluation Metrics

Classification accuracy is a simple and clear indicator of the training and testing performances. However, it does not demonstrate enough details to better understand the performance of the classification model due to increased sensitivity to imbalances among the classes. Hence, we make use of a combination of precision, recall and F1-score over different classes to further compare the performance of all our experimental configurations (Ji, Zhang, & Wu, 2019).

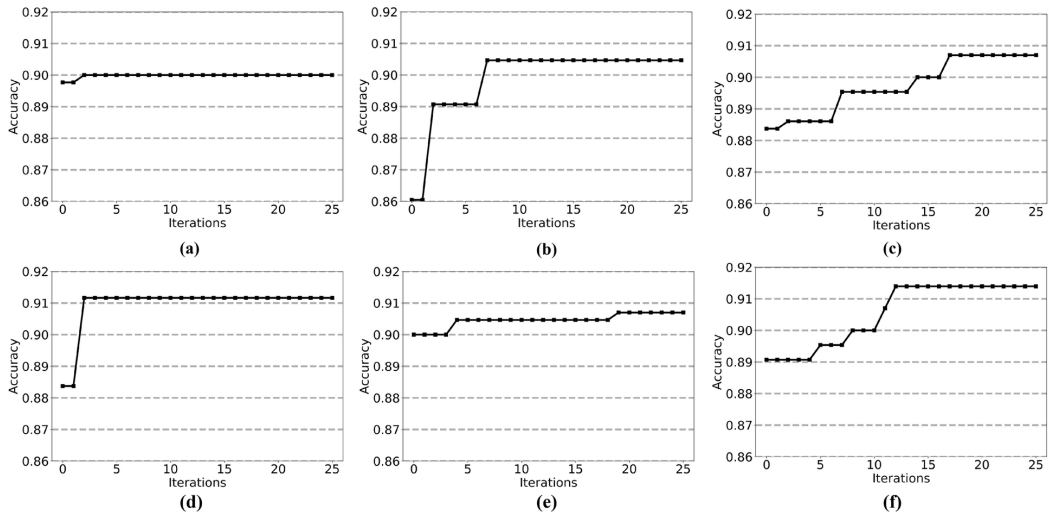
$$precision_c = \frac{TP_c}{TP_c + FP_c} \quad (6)$$

$$recall_c = \frac{TP_c}{TP_c + FN_c} \quad (7)$$

$$F1 - score = \frac{1}{N} \sum_{c=1}^N \frac{2 * recall_c * precision_c}{recall_c + precision_c} \quad (8)$$

where TP is true positive, FP is false positive, FN is false negative and N is the number of diseases types.

Figure 4. Fitness value of training procedures at different hidden layers, number of neurons in hidden layers, activation functions and optimizer functions. (a) PSO-ANN_1 on raw dataset; (b) PSO-ANN_2 on raw dataset; (c) PSO-ANN_3 on raw dataset; (d) PSO-ANN_1 on augmentation dataset; (e) PSO-ANN_2 on augmentation dataset; (f) PSO-ANN_3 on augmentation dataset.



4. RESULTS AND DISCUSSION

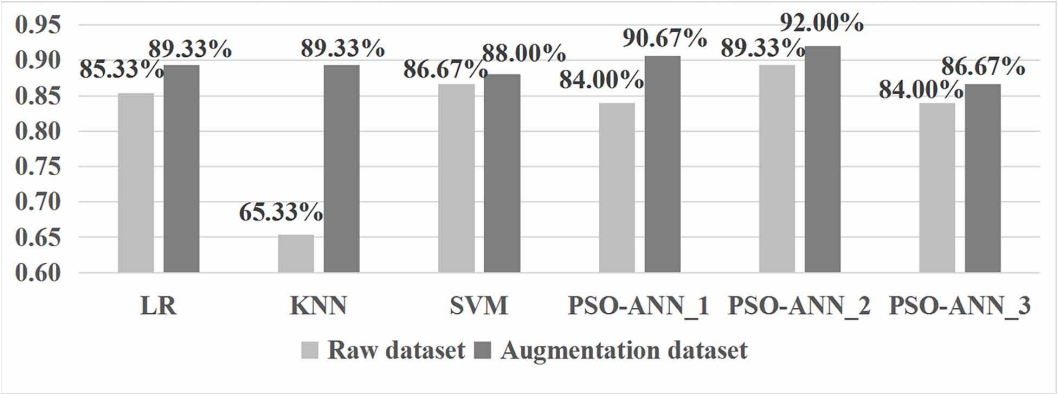
In order to visualize and analyse the training process of PSO-based ANN, the best fitness value in each iteration is recorded. As shown in Figure 6, the searching optimizing process is similar in all PSO-ANN models. Initially, the improvement of the process as the iteration grows and there is no further change on fitness values after multiple iterations. In the end, the fitness values are stable and

show no signs of continuing to increase, showing that PSO-ANN models get sufficiently trained, and the optimized solution can be chosen. In addition, it is not difficult to find that the fewer the number of hidden layers, the faster the model converges. This is because the multi-layer ANNs has greater parameter combinations, so it requires more iterations to determine the optimal model. The results of PSO-based ANN models with optimized parameters combination are demonstrated in Table 2.

Table 2. The results of PSO-based ANN models with optimized parameters combination.

Dataset	Methods	The number of hidden layers	The number of neurons in each hidden layer	Activation	Optimizer
Raw dataset	PSO-ANN_1	1	170	Tanh	Adam
	PSO-ANN_2	2	(174, 138)	ReLU	Adam
	PSO-ANN_3	3	(119, 36, 100)	ReLU	Adam
Augmentation dataset	PSO-ANN_1	1	142	Tanh	Adam
	PSO-ANN_2	2	(63, 61)	ReLU	Adam
	PSO-ANN_3	3	(168, 63, 186)	ReLU	Adam

Figure 5. Test accuracy comparison between different experiment configurations.



In order to measure the performance of all experiment configurations, the accuracy of different classification methods is evaluated on the hold-out test dataset. The results are shown in Figure 5. It is not difficult to find that all models hold the lead in test accuracy on augmentation dataset by a large margin compared with the models trained on raw dataset. It is necessary to carry out dataset augmentation procedure in order to deal with quantitative insufficiency and categorical unbalance of the data. Sufficient and balanced data can avoid over-fitting and thus boost the performance and generalizability of the models. On the other hand, PSO-ANN_2, i.e. PSO-ANN with 2 hidden layers, 63 and 61 neurons in hidden layers respectively, Relu activation function and Adam optimizer achieves the highest test accuracy (92%), though there is no clear advantage in fitness value during validation phase. Consistent with Figure 5, as we can see in Table 3, the average precision (95.00%), recall (92.00%) and F1-score (91.84%) of PSO-ANN_2 are the highest among all these models, which once

again suggests the effectiveness of augmentation dataset and the superiority of the hybrid PSO-ANN model. PSO accelerates the convergence process and improve performance results of ANN via proper parameter selection scheme.

Table 3. Precision, recall and F1-score for the corresponding experimental configurations. The best results are shown in bold.

Methods	Dataset	Raw dataset	Augmentation dataset
	Criterion		
LR	precision	0.8636	0.9320
	recall	0.8533	0.8933
	F1-score	0.8353	0.8837
KNN	precision	0.6725	0.9291
	recall	0.6533	0.8933
	F1-score	0.6035	0.8943
SVM	precision	0.8533	0.8602
	recall	0.8667	0.8800
	F1-score	0.8483	0.8592
PSO-ANN_1	precision	0.8483	0.8810
	recall	0.8400	0.9067
	F1-score	0.8181	0.8852
PSO-ANN_2	precision	0.9370	0.9500
	recall	0.8933	0.9200
	F1-score	0.8953	0.9184
PSO-ANN_3	precision	0.8378	0.8491
	recall	0.8400	0.8667
	F1-score	0.8244	0.8471

The confusion matrices of PSO-ANN_2 model on test dataset are illustrated with an accuracy associated for individual classes which are listed along the diagonal. The confusion matrices are given in terms of percentages, not absolute numbers and the color of the matrix block represents the number of corresponding test samples as shown in the color bar. Figure 6 and Figure 7 demonstrate the results obtained from PSO-ANN_2 model trained on raw dataset and augmentation dataset, respectively. PSO-ANN_2 can distinguish most of diseases types easily, but label8 and label13 are prone to be misclassified possibly due to their similar pathological features. In addition, via data augmentation, PSO-ANN_2 improves the test accuracy of label8 (80% to 100%), label9 (60% to 80%) and label10 (80% to 100%), which contributes a major share to the overall improvement.

Figure 6. Confusion matrices evaluated on test dataset of PSO-ANN_2 trained on raw dataset.

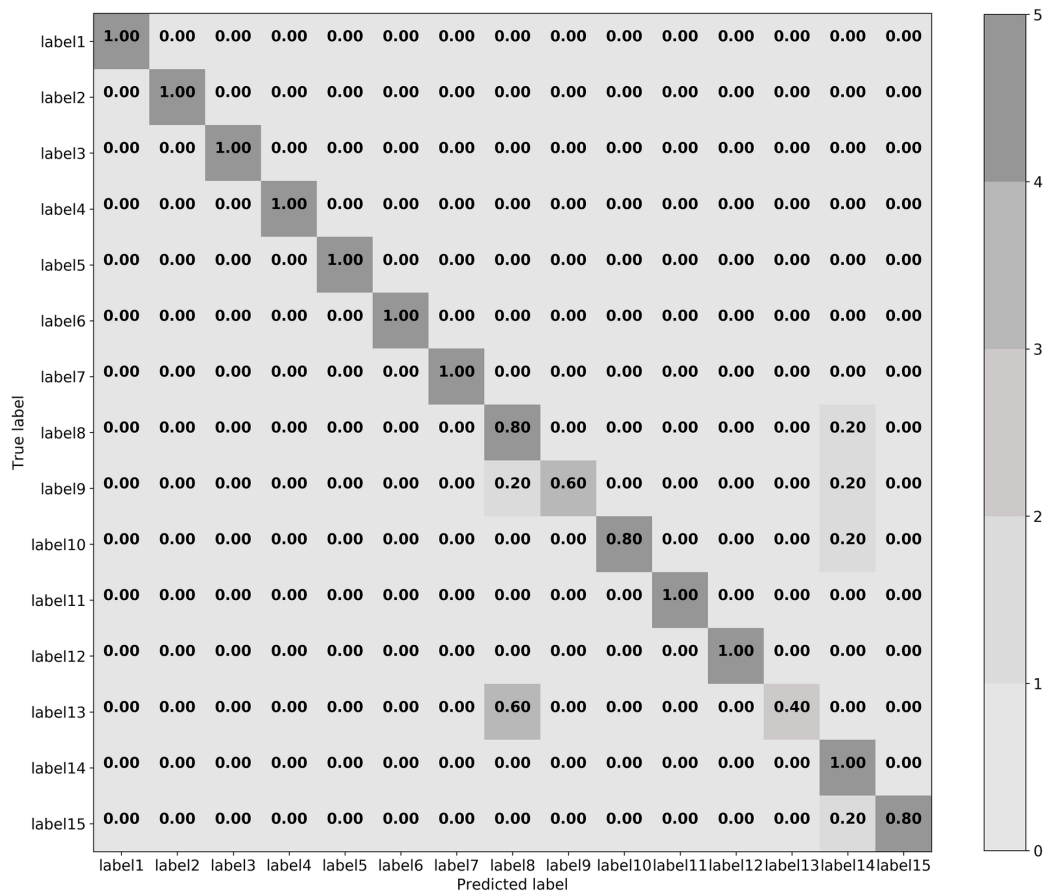
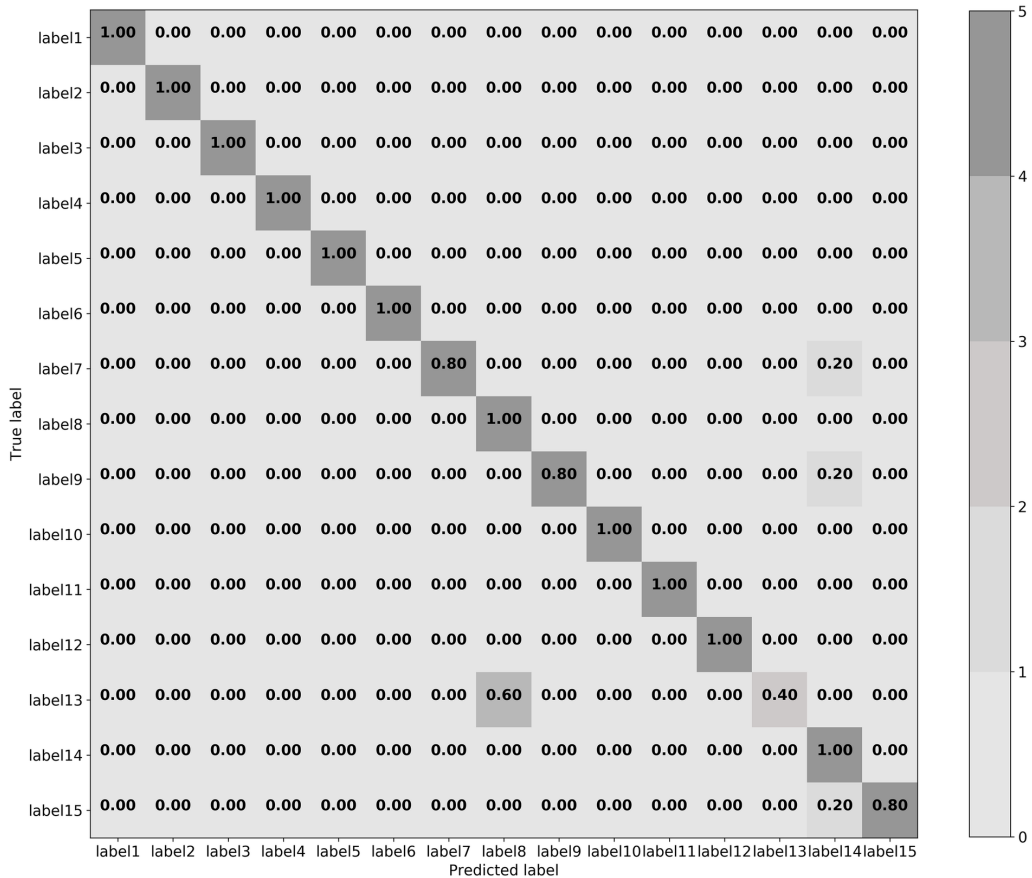


Figure 7. Confusion matrices evaluated on test dataset of PSO-ANN_2 trained on augmentation dataset.



5. CONCLUSIONS

This paper proposes a hybrid PSO-based ANN model (PSO-ANN) for the problem of soybean diseases identification based on the condition of the environment and different features of the soybean, such as the plant stand, the leaves and the seed, etc. PSO algorithm is used to optimize the parameters in ANN, including the activation function, the number of hidden layers, the number of neurons in each hidden layer, and the optimizer. The evaluation of the all experimental configurations is based on a train-validation-test scheme and the performance of each model is assessed on the same hold-out test dataset. Augmentation dataset is created via SMOTE to avoid over-fitting and to boost the performance and generalization of the models. To improve model generalization and avoid randomness, the best parameters in ANN are determined by taking the average validation accuracy of ten times training as the fitness function. In the end, PSO-ANN_2 classification model with 2 hidden layers, 63 and 61 neurons in each hidden layer respectively, activation function ReLu and Adam optimizer, shows higher reliability (92.00%) in soybean diseases identification task compared to the traditional machine learning methods, i.e. LR model (89.33%), KNN model (89.33%), and SVM model (88.00%). Statistician measures of precision, recall, and F1-score as well as confusion matrix are evaluated and also show that PSO can accelerate the convergence process and improve performance results of ANN via proper parameter selection scheme. As a matter of fact, parameters in PSO have a great

influence on the optimization results. In the future, we will focus on the research about PSO algorithm improvement and more advanced optimization methods to update traditional ANNs. Future work will be also focused on state-of-the-art techniques and their application in agricultural area.

In summary, our experiment results show that the proposed PSO-ANN method can be as an effective method for the soybean diseases recognition and can serve as a decision support tool to help farmers to identify the diseases in crops, which is of great significance for modern agriculture development.

ACKNOWLEDGMENT

This work was supported by the Public Welfare Industry (Agriculture) Research Projects Level-2 under Grant 201503116-04-06; Postdoctoral Foundation of Heilongjiang Province under Grant LBHZ15020; Harbin Applied Technology Research and Development Program under Grant 2017RAQXJ096 and National Key Application Research and Development Program in China under Grant 2018YFD0300105-2.

REFERENCES

- Ahmad, J., Muhammad, K., Ahmad, I., Ahmad, W., Smith, M. L., Smith, L. N., Jain, D. K., Wang, H., & Mehmood, I. (2018). Visual features based boosted classification of weeds for real-time selective herbicide sprayer systems. *Computers in Industry*, 98, 23–33. doi:10.1016/j.compind.2018.02.005
- Ahmadi, M.-A., Ahmad, Z., Phung, L. T. K., Kashiwao, T., & Bahadori, A. (2016). Estimation of water content of natural gases using particle swarm optimization method. *Petroleum Science and Technology*, 34(7), 595–600. doi:10.1080/10916466.2016.1153655
- Ahmadi, P., Muharam, F. M., Ahmad, K., Mansor, S., & Abu Seman, I. (2017). Early Detection of Ganoderma Basal Stem Rot of Oil Palms Using Artificial Neural Network Spectral Analysis. *Plant Disease*. 10.1094/PDIS-12-16-1699-RE
- Arslan, M., Güzel, M., Demirci, M., & Ozdemir, S. (2019). *SMOTE and Gaussian Noise Based Sensor Data Augmentation*. Paper presented at the 4th International Conference on Computer Science and Engineering (UBMK). doi:10.1109/UBMK.2019.8907003
- Babu, K. N., Karthikeyan, R., & Punitha, A. (2019). An integrated ANN-PSO approach to optimize the material removal rate and surface roughness of wire cut EDM on INCONEL 750. *Materials Today: Proceedings*, 19, 501–505. Advance online publication. doi:10.1016/j.matpr.2019.07.643
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. doi:10.1613/jair.953
- Da, Y., & Xiurun, G. (2005). An improved PSO-based ANN with simulated annealing technique. *Neurocomputing*, 63, 527–533. doi:10.1016/j.neucom.2004.07.002
- Daniel, P. A., Rafael, S. T., Bruno, S. L., Renata, D. F. L., & Fabyano, R. E. (2017). *Artificial neural network for prediction of the area under the disease progress curve of tomato late blight*. doi:10.1590/1678-992x-2015-0309
- Eberhart, R., & Kennedy, J. (1995). Particle swarm optimization. *Proceedings of the IEEE international conference on neural networks*.
- Ebrahimzade, H., Khayati, G. R., & Schaffie, M. (2019). PSO-ANN-based prediction of cobalt leaching rate from waste lithium-ion batteries. *Journal of Material Cycles and Waste Management*, 1–12. doi:10.1007/s10163-019-00933-2
- Estrada-López, J. J., Castillo-Atoche, A. A., Vázquez-Castillo, J., & Sánchez-Sinencio, E. (2018). Smart Soil Parameters Estimation System Using an Autonomous Wireless Sensor Network With Dynamic Power Management Strategy. *IEEE Sensors Journal*, 18(21), 8913–8923. doi:10.1109/JSEN.2018.2867432
- Hasanipanah, M., Noorian-Bidgoli, M., Armaghani, D. J., & Khamesi, H. (2016). Feasibility of PSO-ANN model for predicting surface settlement caused by tunneling. *Engineering with Computers*, 32(4), 705–715. doi:10.1007/s00366-016-0447-0
- Jamali, B., Rasekh, M., Jamadi, F., Gandomkar, R., & Makiabadi, F. (2019). Using PSO-GA algorithm for training artificial neural network to forecast solar space heating system parameters. *Applied Thermal Engineering*, 147, 647–660. doi:10.1016/j.applthermaleng.2018.10.070
- Ji, M., Zhang, L., & Wu, Q. (2019). Automatic grape leaf diseases identification via UnitedModel based on multiple convolutional neural networks. *Information Processing in Agriculture*. Advance online publication. doi:10.1016/j.inpa.2019.10.003
- Michalski, R. S. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of development an expert system for soybean disease diagnosis. *International Journal of Policy Analysis Information Systems*, 4(2), 125–161.
- Mohamad, E. T., Armaghani, D. J., Momeni, E., Yazdavar, A. H., & Ebrahimi, M. (2018). Rock strength estimation: A PSO-based BP approach. *Neural Computing & Applications*, 30(5), 1635–1646. doi:10.1007/s00521-016-2728-3
- Pan, W. (2012). A new Fruit Fly Optimization Algorithm: Taking the financial distress model as an example. *Knowledge-Based Systems*, 26(26), 69–74. doi:10.1016/j.knosys.2011.07.001

Patil, R., & Tamane, S. C. (2020). PSO-ANN-Based Computer-Aided Diagnosis and Classification of Diabetes. In *Smart Trends in Computing and Communications* (pp. 11-20). Springer.

Roh, S.-B., Oh, S.-K., Park, E.-K., & Choi, W. Z. (2017). Identification of black plastics realized with the aid of Raman spectroscopy and fuzzy radial basis function neural networks classifier. *Journal of Material Cycles and Waste Management*, 19(3), 1093–1105. doi:10.1007/s10163-017-0620-6

Ruiz, L. G. B., Rueda, R., Cullar, M. P., & Pegalajar, M. C. (2018). Energy consumption forecasting based on Elman neural networks with evolutive optimization. *Expert Systems with Applications*, 92, 380–389. doi:10.1016/j.eswa.2017.09.059

Sharma, P., Singh, B., & Singh, R. (2018). *Prediction of Potato Late Blight Disease Based Upon Weather Parameters Using Artificial Neural Network Approach*. Paper presented at the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). doi:10.1109/ICCCNT.2018.8494024

Upendar, J., Gupta, C., Singh, G., & Ramakrishna, G. (2010). PSO and ANN-based fault classification for protective relaying. *IET Generation, Transmission & Distribution*, 4(10), 1197–1212. doi:10.1049/iet-gtd.2009.0488

Wu, Q., Zhang, K., & Meng, J. (2019). Identification of Soybean Leaf Diseases via Deep Learning. *Journal of The Institution of Engineers: Series A*, 100(4), 659–666. doi:10.1007/s40030-019-00390-y

Zhan, Q., Tian, J., & Tian, S. (2019). *Prediction Model of Land Use and Land Cover Changes in Beijing Based on Ann and Markov_CA Model*. Paper presented at the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium.

Miaomiao Ji received her Bachelor's degree in Engineering Management from Xuzhou University of Technology, Xuzhou, China. She is currently a graduate student in Management Science and Engineering at Northeast Agricultural University, Harbin, China. Her current research interests focus on machine learning and intelligent agricultural.

Peng Liu received his Bachelor's degree in Engineering management from Qingdao University of Technology, Qingdao, China. He is currently a graduate student in Management Science and Engineering at Northeast Agricultural University, Harbin, China. His current research interests focus on machine learning and intelligent agriculture.