Research on the identification of the production origin of Angelica dahurica using LIBS technology combined with machine learning algorithms*

SUN Jiaxing[1,2,3], LI Honglian[1,2,3], YAO Yuhang[1,2,3], YAN Qiongyan[1,2,3], and DONG Fang[1,2,3]**

1. College of Quality and Technology Supervising, Hebei University, Baoding 071000, China

2. National and Local Joint Engineering Center of Measuring Instruments and Metrology Systems, Baoding 071000, China

3. Key Laboratory of Energy Measurement and Safety Detection Technology in Hebei Province, Baoding 071000, China

(Received 24 June 2023; Revised 9 August 2023)

The origin of Angelica dahurica medicinal herbs varies, and their pharmacological effects also differ. In order to achieve rapid and accurate identification of the origin of Angelica dahurica medicinal herbs, this study utilizes laser induced breakdown spectroscopy (LIBS) technology combined with machine learning algorithms to identify the original source of Angelica dahurica. Sliced samples of Angelica dahurica were taken from four regions: Hebei, Henan, Zhejiang, and Sichuan. The spectral data from the sliced samples were used as features, and different algorithms including support vector machine (SVM), back propagation (BP) neural network, genetic algorithm-back propagation (GA-BP) neural network, particle swarm optimization-back propagation (PSO-BP) neural network, convolutional neural network (CNN), and CNN-SVM were employed to classify the origin of Angelica dahurica samples. The results show that the average prediction accuracy of the BP, GA-BP, and PSO-BP algorithms reached 89.64%, 89.66%, and 89.93%, respectively. The average prediction accuracy of the SVM, CNN, and CNN-SVM algorithms reached 89.92%, 90.32%, and 90.53%, respectively. The average prediction accuracy improved when the two algorithms were combined, and the CNN-SVM algorithm showed a 44% increase in the lowest prediction accuracy compared to the SVM algorithm. Overall, the combination of the CNN-SVM algorithm and LIBS technology demonstrated the best performance for identifying the origin of Angelica dahurica, a traditional Chinese medicinal herb, and can provide reference for the origin identification of medicinal materials.

Document code: A **Article ID:** 1673-1905(2024)03-0171-6

DOI https://doi.org/10.1007/s11801-024-3114-5

Angelica dahurica is a commonly used Chinese medicinal herb. It has a warm nature and a pungent taste, and is known for its ability to relieve surface wind-cold, dispel wind and alleviate pain, open the nasal passages, dry dampness, eliminate swelling, and promote pus drainage. In daily life, it is frequently used for conditions, such as cold with headache, pain in the eyebrows and temples, toothache, and nasal congestion. The medicinal components of Angelica dahurica are related to its genetic characteristics as well as influenced by the growing environment. Previous studies have indicated that Angelica dahurica from different origins can have varying pharmacological effects[1]. Therefore, identifying the origin of Angelica dahurica is of significant practical importance.

The laser induced breakdown spectroscopy (LIBS) technology is a novel spectroscopic detection technique that can be used for qualitative and quantitative analysis of trace elements in samples. In recent years, LIBS technology has gradually become a fast, accurate, and reliable method for identifying metal elements in drugs or pharmaceutical excipients[2]. In related research in China, ZHENG et al[3] demonstrated the feasibility of using LIBS technology combined with an optimized random forest model to identify the grade of Dendrobium officinale. CAI et al[4] applied LIBS technology in combination with the multiplicative signal correction-improved genetic algorithm-support vector machine (MSC-IGA-SVM) model to trace the origin of Dioscorea opposita Rhizoma slices, and the results showed that using IGA could clearly identify the nonlinear relationships in the spectra and was less affected by noise, with

* This work has been supported by the National Natural Science Foundation of China (No.62173122), the Key Natural Science Projects of Hebei Province (No.F2021201031), and the Funding Project for Introducing Overseas Students in Hebei Province (No.C20210312).

** E-mail: dongfang1023@163.com

the best origin-tracing effect achieved by the MSC-IGA-SVM model. Foreign scholars such as MAGAIHAES et al[5] utilized LIBS technology in conjunction with principal component analysis (PCA) and classification via regression and partial least square regression (CVR-PLSR) models to determine the differences in element concentrations in Citrus fresh leaf seedlings, achieving correct classification rates of 95.10% and 80.96%, respectively. LUKAS et al[6] conducted a comprehensive review of the application of LIBS in recent years and discussed the basic principles, advantages, and limitations of the most commonly used machine learning algorithms.

Although some scholars have conducted research on product classification and identification based on the combination of LIBS technology and machine learning algorithms, there have been few reports on the identification of Angelica dahurica from different origins using LIBS technology. Angelica dahurica from different origins exhibits variations in quality, making it difficult to manually differentiate the origin in the Chinese medicinal herb market. Therefore, this study aims to explore the feasibility of using LIBS technology in conjunction with machine learning algorithms for identifying the origin of Angelica dahurica, providing a reference for rapid differentiation and detection of Chinese medicinal herbs in the future.

The schematic diagram of the LIBS detection system used in the experimental setup is shown in Fig.1. The experimental system is mainly composed of a laser, a reflector, a focusing lens, an electric three-dimensional movable platform, a spectrometer, and a computer. Prior to the experiment, the electric displacement platform needs to be positioned close to the optical fiber, and the position of the fiber probe is adjusted to ensure that the coupling lens fully receives the plasma signal. The laser source is generated from an Nd: YAG pulsed laser (Beijing Reibao Optoelectronics Technology Corporation, wavelength: 1 064 nm, Dawa 200). The laser emitted by the laser source follows the optical path shown in Fig.1, reaches the transmission mirror bracket, and is reflected by a 90° mirror to the focusing lens. The focused laser beam passes through the empty tube and hits the sample surface, thereby exciting the plasma. After the fiber probe detects the plasma, the data is transmitted to the spectrometer (Ocean Optics Corporation, USA, MAX2500+), and the processed LIBS spectrum is displayed on the computer[7].

The Angelica dahurica herbal medicine used in this study was sourced from four different regions: Hebei, Henan, Zhejiang, and Sichuan. Firstly, sample pretreatment was performed by cleaning and air-drying the Angelica dahurica samples. Subsequently, the air-dried Angelica dahurica was placed in a grinder for pulverization. The pulverized samples were then sieved using a 100-mesh sieve to remove larger particles. The sieved samples were further ground in a mortar and pestle for 10 min. Finally, five samples were taken from each region, with each sample weighing 2 g. The 796YP-15A powder press machine with a pressure of 18 MPa was used to compress the samples for 10 min. The resulting 20 experimental samples were individually wrapped in weighing paper to prevent abrasion and damage between samples. Fig.2 illustrates the process of sample preparation.

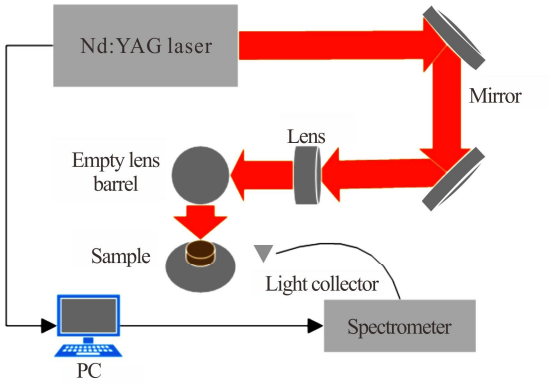

Fig.1 Schematic diagram of the LIBS experimental setup

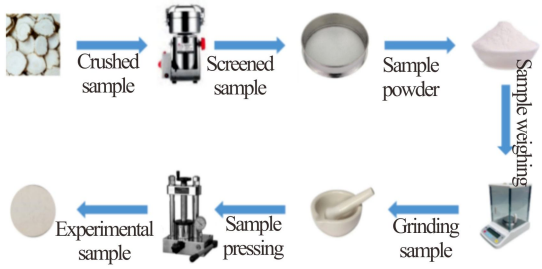

Fig.2 Sample production flow chart

Angelica dahurica from Hebei province contains trace elements such as Fe, Ca, Zn, Cu, Na, and Mn. The characteristic spectral lines of these elements were identified by referring to the NIST database and marked as shown in Fig.3. The positions of the characteristic spectral lines are as follows: Fe II 238.1 nm, 259.8 nm, 393.2 nm; Fe I 422.6 nm; Mn II 247.9 nm, 396.7 nm; Mn I 279.4 nm, 445.3 nm; Zn II 280.1 nm; Cu II 285.1 nm; Na II 317.9 nm. In this experiment, a laser energy of 50 mJ was used to acquire the full spectrum graph within the range of 199—517 nm.

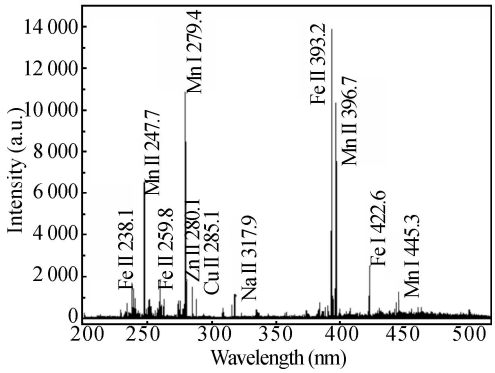

Fig.3 LIBS spectra of Angelica dahurica samples

The samples used for origin identification research are Angelica dahurica samples from four different regions. Five parallel samples were produced for each region, resulting in a total of 20 experimental samples. For each sample, 10 sets of spectral data were collected. Each data set consisted of the average of 7 accumulated laser pulses. In total, 200 spectral data were obtained. The average LIBS spectra of Angelica dahurica samples from different regions are shown in Fig.4. From the figure, it can be observed that the elemental composition of the samples from the four regions is consistent, with variations in the spectral intensities of the individual element lines. However, it is not possible to accurately determine the origin of the Angelica dahurica samples solely based on the spectral graphs. Therefore, machine learning algorithms were utilized to analyze the spectral data and predict the origin of Angelica dahurica.

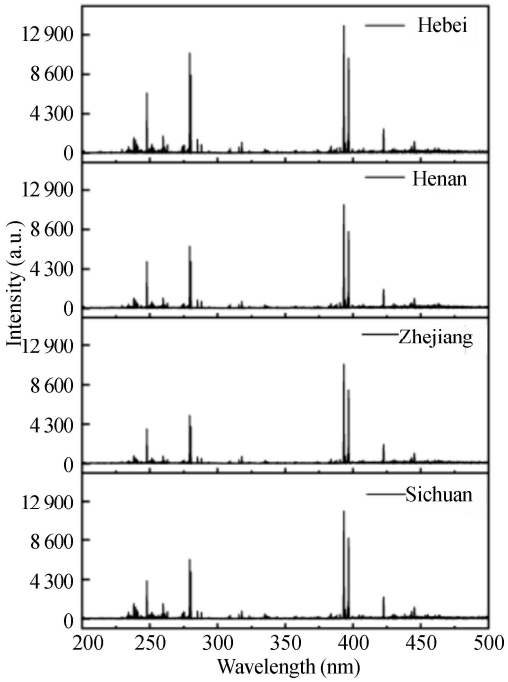

Fig.4 Average LIBS spectra of Angelica dahurica samples from different origins

Seven characteristic spectral lines were selected as features for Angelica dahurica, and their corresponding wavelengths are shown in Tab.1. The selection of the training and testing sets followed a 3: 1 split principle. Approximately 75% of the spectral data from each origin were randomly chosen as the training set, while the remaining 25% were used as the testing set. To reduce experimental errors, each machine learning algorithm model was run 500 times, and the resulting accuracy results were plotted as a distribution graph. The average predicted accuracy was then taken as the evaluation criterion for the model's predictive performance.

The support vector machine (SVM) is a supervised learning method that has the advantage of being able to select parameters and obtain the optimal classification solution even with small sample sizes[8]. In this experiment, a non-linear SVM was used for classification. Firstly, grid search was used to optimize the hyperparameters C and γ of the SVM model. The SVM model was run 500 times, and the prediction accuracy is shown in Fig.5. When the predicted origin overlaps with the true origin on the graph, it indicates accurate prediction. From the graph, it can be observed that the SVM model shows good predictive performance, demonstrating its effectiveness in the origin identification of Angelica dahurica.

Tab.1 Characteristic spectral lines and corresponding wavelengths of Angelica dahurica

Element	Wavelength (nm)
Mn II	247.7, 396.7
Mn I	279.4
Zn II	280.1
Na II	317.9
Fe II	393.2
Fe I	422.6

The accuracy obtained from 500 independent experiments is fitted to a Gaussian curve, as shown in Fig.6. The average accuracy for predicting the origin of Angelica dahurica is 89.92%, with the lowest accuracy being 34% and the highest being 100%. Due to the sensitivity of SVM to features in input data, the presence of noise, outliers, or inconsistencies in the data can cause significant fluctuations in the prediction accuracy of SVM. Therefore, when LIBS technology is combined with SVM for identifying the origin of Chuanxiong, the results may exhibit randomness.

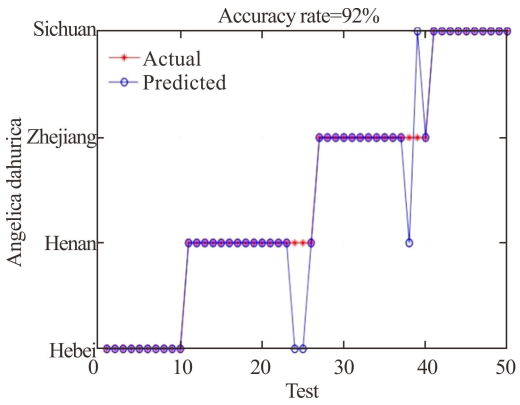

Fig.5 Identification map of SVM

The back propagation (BP) neural network is a multi-layer feedforward network that utilizes error backward propagation[9]. The BP neural network is trained using the training set to determine the weights and thresholds that minimize the error. However, it has the disadvantage of slow learning convergence and may not guarantee convergence to the global minimum point.

Genetic algorithm (GA) and particle swarm optimization (PSO) algorithms possess excellent global characteristics, enabling them to quickly find the optimal weights and thresholds for the BP neural network. This helps to prevent premature convergence of the network and improves the accuracy of classification[10]. Therefore, GA and PSO algorithms are employed to optimize the BP neural network.

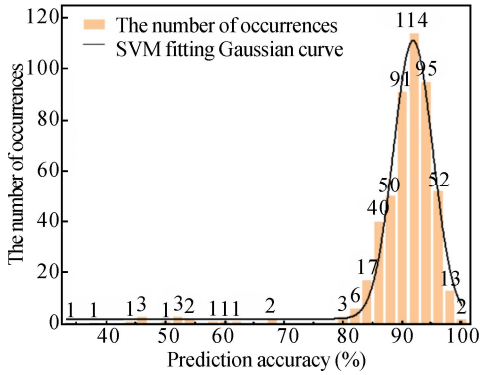

Fig.6 Results of SVM after 500 cycles

The training parameters are set as follows: maximum iteration number of 1 000, target training error of 10^{-6}, and learning rate of 0.01. Since the data features consist of 7 elements, the input layer has 7 nodes, the hidden layer has 6 nodes, and the output layer has 4 nodes. In GA-BP, the initial optimization parameter genetic generation is set to 50, and the population size is set to 5. After initializing the population and optimizing the algorithm, the optimal parameters are obtained. These parameters are then used to train the BP model and classify the test set. In PSO-BP, the population, velocity, individual best, and global best are iteratively optimized to determine the optimal initial weights and thresholds. These optimized values are then used to classify the test set.

Fig.7 illustrates the comparison of prediction accuracies for BP, GA-BP, and PSO-BP neural network algorithms after 500 cycles. The average prediction accuracy for BP is 89.64%, while the average prediction accuracies increase to 89.66% and 89.93% after optimizing BP using GA and PSO algorithms, respectively. The GA-BP algorithm can be used to simultaneously optimize the weights and structural parameters of a neural network. The PSO-BP algorithm, on the other hand, considers the simultaneous optimization of weights and biases. Compared to the GA-BP algorithm, the PSO-BP algorithm has stronger global search capabilities. It does not require the selection of a fitness function or adjustment of parameters related to genetic algorithms. The fine adjustment of weights and biases in the PSO-BP algorithm makes it easier to find optimized solutions that fit the characteristics of the problem. Therefore, the PSO-BP algorithm has better predictive performance. Tab.2 provides a comparison of prediction accuracies for the three neural network algorithms.

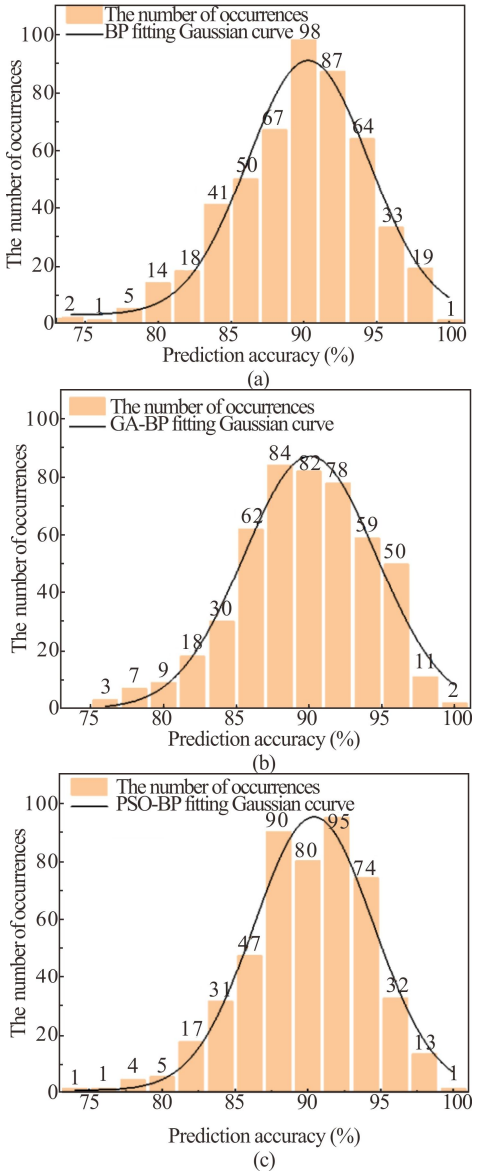

Fig.7 Results comparison of (a) BP, (b) GA-BP and (c) PSO-BP algorithms after 500 cycles

Tab.2 Prediction accuracy of BP, GA-BP, and PSO-BP algorithms after 500 cycles

Algorithm	Minimum prediction accuracy (%)	Highest prediction accuracy (%)	Average prediction accuracy (%)
BP	74	100	89.64
GA-BP	76	100	89.66
PSO-BP	74	100	89.93

The convolutional neural network (CNN) algorithms belong to deep learning and are widely used in the field. CNN algorithms exhibit stronger performance in feature

extraction compared to manual methods, but have slower convergence and are prone to overfitting[11]. To prevent overfitting, set the number of CNN model layers to 1 and use pooling layers before output. Batch normalization is applied to accelerate the convergence speed. The constructed CNN model consists of an input layer for data input, followed by two convolutional layers, batch normalization layers, ReLU activation layers, and pooling layers. The convolutional layers have a kernel size of 2×1, and the pooling layers use the maximum pooling method with a window size of 2×1 and a stride of 2. Finally, the classification is performed after the normalization layer with the number of categories set to 4. The CNN network uses the Adam gradient descent algorithm with the following parameter settings: maximum training iterations of 500, batch size of 128, and learning rate of 0.01.

The accuracy obtained from the CNN algorithm after 500 cycles is fitted to a Gaussian curve as shown in Fig.8. The experimental results show that the average prediction accuracy for identifying the origin of Angelica dahurica is 90.32%. The lowest accuracy achieved is 76%, while the highest accuracy is 100%. The CNN algorithm has less fluctuation in the prediction accuracy for the origin identification of Angelica dahurica. Due to the ability of CNN models to be pre-trained on large-scale datasets, they can learn generalized feature representations. Then, by fine-tuning the model to adapt it to the task of identifying the origin of Chinese medicinal herbs, training time can be saved and identification accuracy can be improved.

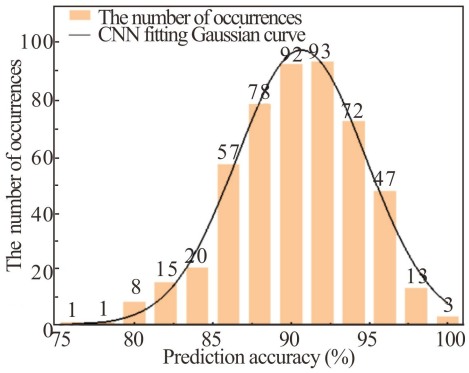

Fig.8 Results of CNN after 500 cycles

Regarding the CNN algorithm, although it excels in feature extraction, it tends to converge slowly. As the number of iterations increases, it is more suitable for linearly separable cases. The optimal hyperplane constructed by the multi-layer perceptron classifier[12] is not necessarily the optimal solution[13-15]. On the other hand, SVM is capable of mapping nonlinear problems into linear ones[16] and has the advantage of handling small sample sizes. It is suitable for the small sample size of Angelica dahurica in this study and can address the issue of overfitting. In this study, the CNN model is used for feature extraction from spectral data, and the SVM model is used for classification and prediction.

The parameters for the CNN-SVM model in this study are as follows: the maximum training iterations are set to 500, the initial learning rate is 0.001, and the dataset is shuffled before each training. The training results at a prediction accuracy of 96% are shown in Fig.9. From the figure, it can be observed that the accuracy increases and converges as the number of iterations increases, and the loss function decreases and tends towards zero.

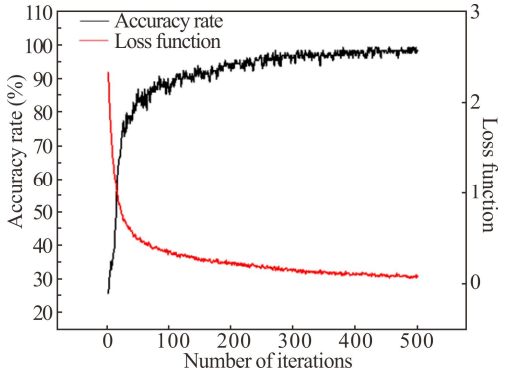

Fig.9 Results of CNN-SVM

The accuracy obtained from 500 independent experiments of the CNN-SVM algorithm is fitted with a Gaussian curve, as shown in Fig.10. The average accuracy for predicting the origin is 90.53%, with a minimum of 78% and a maximum of 100%. The average accuracy of origin prediction was 90.53%, with the lowest being 78% and the highest being 100%. Compared to the SVM and CNN algorithms, the lowest prediction accuracy was improved by 44% and 2%, respectively. The overall fluctuation in accuracy over 500 predictions was minimal, and the average prediction accuracy was higher than that of SVM and CNN algorithms. The CNN-SVM model can consider the relationships between samples during the training process. SVM acts as a linear classifier, separating samples from different classes with an optimal decision boundary. On the other hand, CNN extracts local and global features through convolutional layers while preserving relative spatial information. By considering local and global features as well as sample relationships, it can better capture the differences between samples and improve the accuracy of origin identification. Tab.3 presents a comparison of the prediction accuracies of the SVM, CNN, and CNN-SVM algorithms.

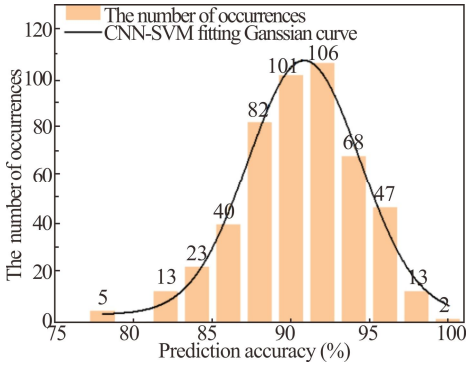

Fig.10 Results of CNN-SVM after 500 cycles

Tab.3 Prediction accuracy of SVM, CNN and CNN-SVM algorithms after 500 cycles

Algorithm	Minimum prediction accuracy (%)	Highest prediction accuracy (%)	Average prediction accuracy (%)
SVM	34	100	89.92
CNN	76	100	90.32
CNN-SVM	78	100	90.53

In this study, the LIBS technique was adopted to detect the traditional Chinese medicinal herb Angelica dahurica, revealing its rich content of trace elements such as Fe, Ca, Zn, Cu, Na, and Mn. By combining machine learning algorithms, the origin of Angelica dahurica from Hebei, Henan, Zhejiang, and Sichuan was distinguished. The experimental results showed that the SVM algorithm achieved an average prediction accuracy of 89.92%. The average prediction accuracies of the neural network models, including BP, GA-BP, PSO-BP, and CNN, were 89.64%, 89.66%, 89.93%, and 90.32%, respectively. The CNN-SVM model had an average prediction accuracy of 90.53%. Moreover, the CNN-SVM model improved the lowest prediction accuracy of the SVM algorithm by 44%, indicating a significant enhancement in performance. Additionally, the overall fluctuation in prediction accuracy was minimal for CNN-SVM. The experimental results demonstrated that the CNN-SVM model had the best predictive performance and effectively improved the accuracy of Angelica dahurica origin identification. The findings of this study provide reference for the identification of Angelica dahurica origin based on the LIBS technique. In the future, further research can be conducted on unsupervised learning methods and real-time origin identification.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] WANG L, SUN J, CHEN M, et al. Genetic diversity and quality traits of Angelica dahurica from different production regions[J]. Journal of Zhejiang University, 2023, 40(01): 30-37. (in Chinese)

[2] WU Y, CAO L, WANG Y, et al. Identification of metal elements in Chinese medicinal materials or excipients using LIBS spectroscopy[J]. Journal of pharmaceutical analysis, 2019, 39(03): 557-564. (in Chinese)

[3] ZHENG P C, ZENG S, WANG J M, et al. Study on recognition of dendrobium officinale grades using LIBS[J]. Spectroscopy and spectral analysis, 2020, 40(03): 941-944. (in Chinese)

[4] CAI Y, ZHAO Z F, GUO L B, et al. Traceability study of dioscorea opposita herbal slices based on LIBS[J]. Spectroscopy and spectral analysis, 2023, 43(01): 138-144. (in Chinese)

[5] MAGAIHAES A B, SENSI G S, RANULFI A, et al. Discrimination of genetically very close accessions of sweet orange (citrus sinensis L. Osbeck) by laser-induced breakdown spectroscopy (LIBS)[J]. Molecules, 2021, 26(11): 3092.

[6] LUKAS B, ZUZANA G, HANS L, et al. A critical review of recent trends in sample classification using laser-induced breakdown spectroscopy (LIBS)[J]. Trends in analytical chemistry, 2022: 116859.

[7] SUN J X, LI H L, LV H S, et al. Research on heavy metal detection based on laser-induced breakdown spectroscopy technology under magnetic field constraint[J]. Journal of optoelectronics·laser, 2023, 34(04): 422-428. (in Chinese)

[8] JAKUB N, MICKAL K. Selecting training sets for support vector machines: a review[J]. Artificial intelligence review, 2019, 52(2): 857-900.

[9] WU C T, WU L X, QIU C H, et al. Experimental and numerical studies on lithium-ion battery heat generation behaviors[J]. Energy reports, 2023, 9: 5064-5074.

[10] GAO H, XUE L Y. Fitting LED spectral model with back propagation neural network based on improved genetic algorithm[J]. Progress in laser and optoelectronics, 2017, 54(07): 294-302. (in Chinese)

[11] HAN Q, YIN C, DENG Y Y, et al. Towards classification of architectural styles of Chinese traditional settlements using deep learning: a dataset, a new framework, and its interpretability[J]. Remote sensing, 2022, 14(20): 5250.

[12] DANA B H, MOHAMMAD K. A recursive general regression neural network (R-GRNN) oracle for classification problems[J]. Expert systems with applications, 2019, 135: 273-286.

[13] WANG Y, YANG L. Joint learning adaptive metric and optimal classification hyperplane[J]. Neural networks, 2022, 148: 111-120.

[14] WANG Y, HONG K, ZOU J, et al. A CNN-based visual sorting system with cloud-edge computing for flexible manufacturing systems[J]. IEEE transactions on industrial informatics, 2019, 16(7): 4726-4735.

[15] JIRAPONG M, LUISE P, ACHIM S, et al. Human forehead recognition: a novel biometric modality based on near-infrared laser backscattering feature image using deep transfer learning[J]. IET biometrics, 2020, 9(1): 31-37.

[16] SUN B, JIANG D, ZUO Z, et al. Gender recognition via fused silhouette features based on visual sensors[J]. IEEE sensors journal, 2019, 19(20): 9496-9503.