scientific reports

Check for updates

OPEN

Rapidly detecting fennel origin of the near-infrared spectroscopy based on extreme learning machine

Enguang Zuo[1,5], Lei Sun[4,5], Junyi Yan[2], Cheng Chen[1,2]✉, Chen Chen[2]✉ & Xiaoyi Lv[1,2,3]

Fennel contains many antioxidant and antibacterial substances, and it has very important applications in food flavoring and other fields. The kinds and contents of chemical substances in fennel vary from region to region, which can affect the taste and efficacy of the fennel and its derivatives. Therefore, it is of great significance to accurately classify the origin of the fennel. Recently, origin detection methods based on deep networks have shown promising results. However, the existing methods spend a relatively large time cost, a drawback that is fatal for large amounts of data in practical application scenarios. To overcome this limitation, we explore an origin detection method that guarantees faster detection with classification accuracy. This research is the first to use the machine learning algorithm combined with the Fourier transform-near infrared (FT-NIR) spectroscopy to realize the classification and identification of the origin of the fennel. In this experiment, we used Rubberband baseline correction on the FT-NIR spectral data of fennel (Yumen, Gansu and Turpan, Xinjiang), using principal component analysis (PCA) for data dimensionality reduction, and selecting extreme learning machine (ELM), Convolutional Neural Network (CNN), recurrent neural network (RNN), Transformer, generative adversarial networks (GAN) and back propagation neural network (BPNN) classification model of the company realizes the classification of the sample origin. The experimental results show that the classification accuracy of ELM, RNN, Transformer, GAN and BPNN models are above 96%, and the ELM model using the hardlim as the activation function has the best classification effect, with an average accuracy of 100% and a fast classification speed. The average time of 30 experiments is 0.05 s. This research shows the potential of the machine learning algorithm combined with the FT-NIR spectra in the field of food production area classification, and provides an effective means for realizing rapid detection of the food production area, so as to merchants from selling shoddy products as good ones and seeking illegal profits.

As one of the popular spices that are widely used as condiments in daily life[1], the fennel is widely planted all over the world[2]. In addition, the fennel can also be used as a raw material for the production of wine, creams, perfumes, biochemical materials, etc.[3–5], and has some medicinal value. It can be used to prepare many drugs[6], such as in the breeding industry and other fields. The antibacterial properties of fennel can replace antibiotics in feed additives and medicines in the poultry industry, and can effectively prevent the abuse of antibiotics[7]. On the other hand, the fennel and its derivatives have many beneficial medical properties, which can be used to treat digestive diseases such as epilepsy, anorexia and abdominal distension[8], lower blood pressure, antibacterial and anti-inflammatory, prevent cancer, and treat diabetes[6,9,10].

In recent years, a large number of researchers have focused on the chemical composition of the fennel because of its versatility, and found that the efficacy of the fennel and its products is related to the type and amount of the chemicals contained, such as antioxidant phenols and antibacterial terpenes. In addition, studies have shown that the relative concentration of compounds contained in fennel largely depends on its geographical origin[11], because the types and content of biochemical components of fennel are origin dependent on geographical location[12]. Due

[1]College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China. [2]College of Software, Xinjiang University, Urumqi 830046, China. [3]Key Laboratory of signal detection and processing, Xinjiang University, Urumqi 830046, China. [4]Xinjiang Uygur Autonomous Region Product Quality Supervision and Inspection Research Institute, Urumqi 830011, China. [5]These authors contributed equally: Enguang Zuo and Lei Sun. ✉email: chenchengoptics@gmail.com; 1343432873@qq.com

to the influence of the external environment such as the climate, temperature, soil conditions, and precipitation of the production area, and the combined effect of internal and external factors such as genotypes[13,14], the kinds and contents of substances and yields of fennel grown in different regions are significantly different[12,15–17]. The main biologically active ingredients contained in the fennel, such as antibacterial microorganisms, phenols and fatty acids, are different in content[12], resulting in different quality, nutritional content and prices[16]. Yingdong College of Biological Engineering, Shaoguan University, Guangdong Province, China compared the nutrient content of 7 fennel varieties from Yumen, Gansu and Yili, Xinjiang, and comprehensive evaluation showed that the nutritional value of fennel varieties from Yumen, Gansu was relatively high. In order to prevent some merchants from using shoddy products as good ones and causing obstacles to market supervision, it is of great significance to accurately identify the origin of the fennel. However, the existing studies are mainly focused on the analysis of the differences in the kinds and contents of chemical substances in fennel from different regions[12,17], and there are still gaps in the classification and identification of the origin of the fennel.

Different from the traditional food production area classification methods, such as molecular biotechnology, multi-element and multi-isotope analysis, which have complicated steps and high detection requirements[18,19], the FT-NIR spectroscopy has the advantages of simple sample preparation and low detection cost[20] In recent years, the FT-NIR spectroscopy combined with machine learning algorithms have been widely used in the field of food production area classification. Many research teams have completed the origin classification of tea, pistachio fruit, olive oil, multiflorum, cocoa bean, wheat, honey, radix glycyrrhizae and other foods by the FT-NIR combined with the machine learning algorithm[21–29]. As a nonlinear neural network, the BPNN (back propagation neural network) can solve complex problems more accurately than linear neural networks[28,30]. Xiu Ying Liang et al. used the BPNN combined with the FT-NIR spectroscopy to classify honey from different flower lines, the accuracy could reach 100% in a specific spectral segment[27]. Yan Tian ying et al. used the convolutional neural network (CNN) and recurrent neural network (RNN) models combined with the NIR spectroscopy to identify the geographical origins of Radix Glycyrrhizae from Gansu, Inner Mongolia, Ningxia, and Xinjiang, respectively, with the classification accuracy could reach 93%[29]. Yang et al. proved that when generative adversarial network (GAN) has a small amount of training data, the input data does not need feature selection, but also the model obtained by competitive learning is better than other classification algorithms, which provides a new method for the existing infrared spectrum research[28]. However, the time cost of GAN in the classification process is relatively high. For practical application scenarios, the detection data in the food production area classification is usually massive, which puts a higher demand on the detection time.

Compared with the traditional neural networks, ELM (extreme learning machine) has several remarkable characteristics: easy to use, faster-learning speed and higher generalization performance[31]. In short, the ELM-based model can be faster than the traditional learning algorithm on the premise of ensuring accuracy, so it is more suitable for the rapid detection of food production. Felix YH Kutsanedzie et al. used the ELM combined with the NIR spectroscopy to complete the classification of three grades of cocoa beans, with an accuracy of 94%[32]; Wenbin Zheng et al. found that the performance of the ELM was much better than that of the KNN, LS-SVM and BPNN in the application of the NIR spectroscopy for food classification, which indicates that the ELM may be a promising real-time food classification method[33]. Therefore, in this study, we aim to realize the rapid identification of fennel from different origins by the FT-NIR spectroscopy combined with the ELM and deep learning models, so as to provide an intelligent supervision method for the phenomenon of different the good and bad products in the fennel market caused by some merchants selling shoddy products.

Materials and method

Sample preparation and plant statement. The fennel seeds used in this experiment were harvested in 2020 and we purchased them in batches from local spice companies in Turpan, Xinjiang and Yumen, Gansu in different seasons to ensure that the experimental samples were from the origin and that the effects of different climatic conditions on the samples were taken into account. 200 samples were collected from Lianyungang Kaihao Tong Trading Co. in Xinjiang, and 116 samples were collected from Gansu Yumen Xiaosannong Taobao online store. Since the fennel seeds contain a variety of volatile components, such as essential oils, heating and drying should be avoided when storing the samples[34]. We stored the prepared samples in a dry and airtight atmosphere at room temperature for one week, then put them into the pulverizer. The sample powder was passed through a 200-mesh sieve and subsequently stored in a sealed self-sealing bag. It should be noted that we comply with relevant institutional, national, and international guidelines and legislation of this paper collected experiment sample.

Spectral data collection and preprocessing. All samples in this experiment were measured indoors at a room temperature of 22°C. Before each measurement, OPUS 65 software was used to measure the atmospheric background data under the Windows XP system environment, and then the background data of each sample was measured. Vertex 70 FT-NIR spectrometer (Bruker) was used in the experiment. CO_2 compensation was selected as the atmospheric compensation parameter. The scanning parameters were as follows: the scanning range was 4000–11000 cm^{-1}, the resolution was 8 cm^{-1}, and the scanning was repeated 32 times. The spectrum data of 116 cases of fennel seed samples in Yumen, Gansu, and 200 cases of fennel seed samples in Turpan, Xinjiang were obtained. A total of 316 cases of FT-NIR spectrum data were obtained.

The original FT-NIR spectrum collected by the instrument contains many interference factors, which will affect the classification effect of the model[35,36]. Spectral preprocessing is to reduce or eliminate the influence of interference factors on the spectrum, thus improving the accuracy and reliability of the classification model. Baseline correction is one of the most commonly used methods in preprocessing. The experiment uses the Rubberband baseline correction method, and the baseline point value is 64. In addition, the environmental factors

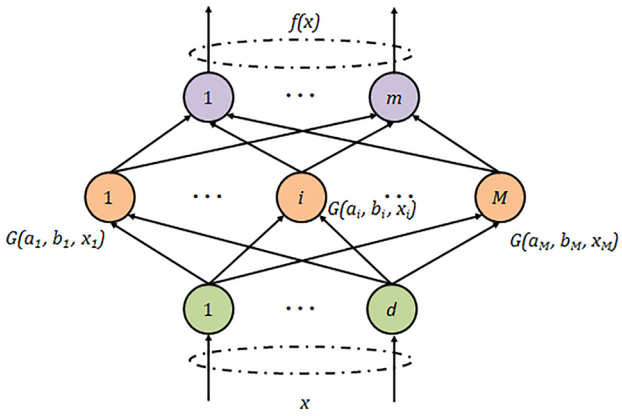

Figure 1. Neural network structure of ELM.

during spectral data collection and preprocessing were strictly consistent to ensure that environmental differences between samples did not interfere with the detection of fennel origin, which is consistent with many current studies based on FT-NIR spectral data to distinguish the geographical origin of plant foods[37,38].

Feature extraction. Principal Component Analysis (PCA) is a classic unsupervised feature extraction algorithm that can reduce the dimensionality of the data[39]. PCA compresses the highly correlated original variables into a few new variables through linear transformation, seeking new variables (that is, principal components) that can maximize the data structure characteristics of the original variables. The first variable has the largest variance and becomes the first principal component, followed by the second variable, and so on.

In this study, 221 samples were randomly selected from 316 samples as the training set and 95 samples as the test set in a ratio of 7:3 for each trial. The cross-validation method was used to randomly divide the data set into training and validation sets, and the randomly generated subsamples were repeatedly applied for training and validation. The PCA algorithm (MATLAB R2019b) was used to extract the features of Rubberband baseline correction spectral data.

Classification model. As a feedforward neural network, the ELM (Extreme Learning Machine) has shown good performance in the application of dataset classification in recent years[31]. The neural network structure of the ELM model is shown in Fig. 1, and the inputs and outputs of the ELM are x and f(x), respectively. The sampled sample data is x_i, and the activation function is G(a_j,b_j,x), where a_j and b_j denote the connection weights and bias values between the input and hidden layers, respectively. Unlike traditional neural network learning algorithms, the ELM not only tends to reach the minimum training error but also the minimum output weight norm. Bartlett's theory shows that for feedforward neural networks if the training error is minor and the weight norm is smaller, the network tends to have better generalization performance[40]. Significantly, ELM has a faster learning speed with great performance; thus, we choose ELM as the first kind of classification model, the activation function kinds are set to Sigmoid, Sine, Hardlim, and the number of neurons in the hidden layer is set to 100. Three models are obtained: ELM-sig, ELM-sin, ELM-hardlim.

The BPNN algorithm is one of the most widely used neural network models, which is a multi-layer feedforward network trained by error backpropagation. BPNN has good ability for self-learning, self-adaptation and generalization, and has excellent effects when used in binary classification[41]. In recent years, BPNN has been widely used in food research, biomedicine and other fields[42–44]. In this study, we choose BPNN as the second kind of classification model, set tansig and logsig as the activation function of the hidden layer, and the output layer function is purelin. The training function of the neural network is trainlm. Set the number of neural network training to 500 times, the learning rate parameter to 0.01, and the learning target parameter to 0.1. Two models are obtained: BPNN-tansig and BPNN-logsig.

The RNN introduces state variables to store past information, and uses state variables with the current input to determine the current output. LSTM is a typical variant of RNN, it is particularly suitable for the classification of sequential data. In recent years, LSTM has been widely used in biomedical classification task[45]. We choose RNN as the third classification model. We set the number of hidden layers to 3, the probability of the dropout layer to 0.5, the learning rate to 0.0001, the batch size to 8 and the epoch size to 20.

The transformer network is developed based on the attention mechanism, which consists of an encoder and decoder for handling long-term dependencies in sequence-to-sequence tasks. In addition, it has excelled in classification tasks[46]. Therefore, we choose it as the fourth classification model. We apply dropout to the output of each sublayer, which is then added to the input of the sublayer and normalized, where the dropout is set to 0.5. Set the learning rate to 0.00001, the batch size to 8 and the epoch size to 40.

The CNN has two features of sparse connectivity and weight sharing, so it is less computational than multilayer perceptron (MLP) and its performance is better. In the field of spectroscopy, the one-dimensional (1D) CNN models are used to learn and predict spectra with good performance[47]. In this study, it is used as the fifth

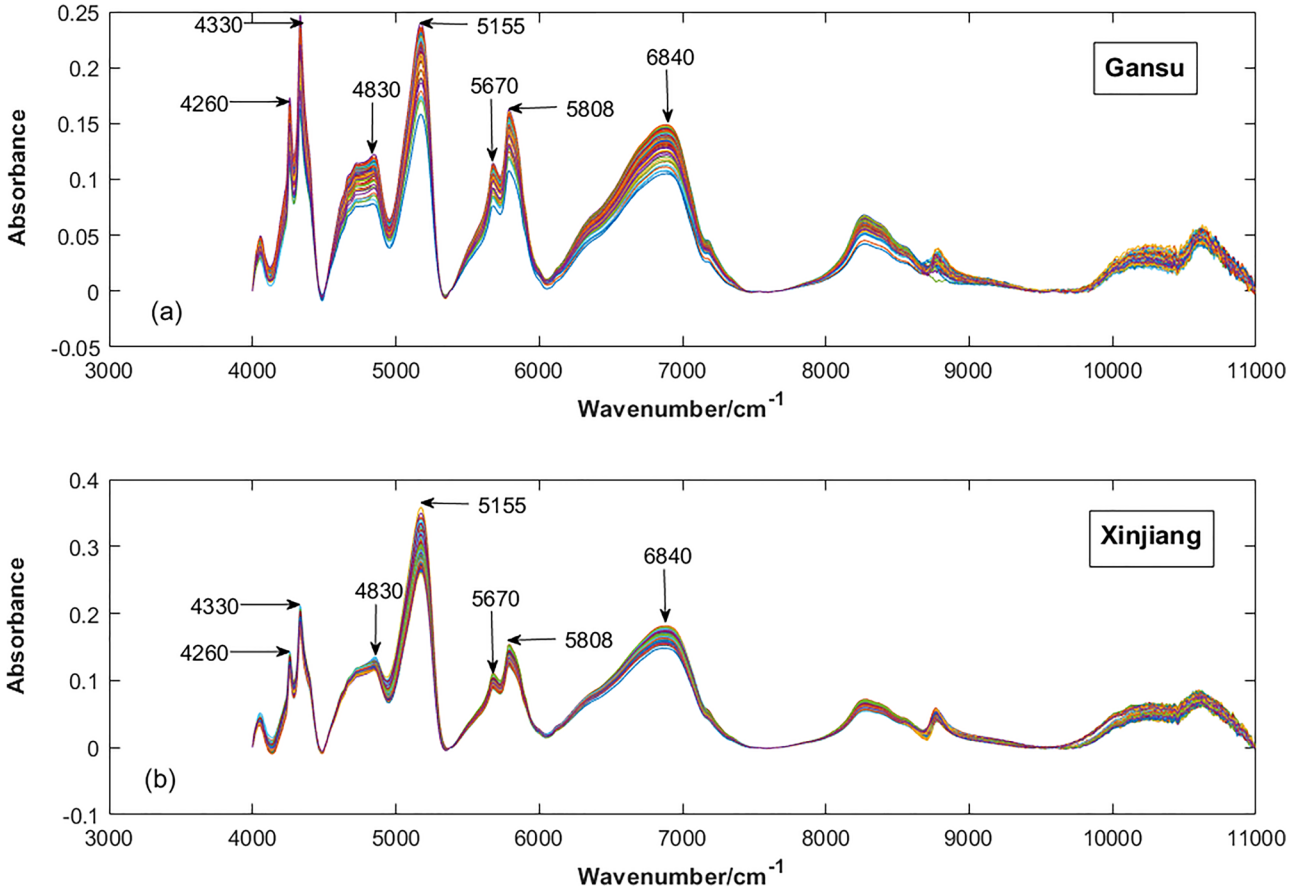

Figure 2. The NIR spectrum of the fennel.

classification model. We choose ReLU as the activation function and set the learning rate to 0.001, the probability of the dropout layer to 0.5, the batch size to 16 and the epoch size to 75.

GAN consists of two artificial neural networks, G and D. The two networks compete with each other to achieve improvement in model performance during training. GAN shows high accuracy in food detection and classification[28]. So we choose it as the sixth classification model. In the fennel classification experiments, we need to train four models from two origins, including G1, G2, D1 and D2. We use G to generate the false spectral data, and then use D to determine the generated data is the probability of being true, and the next step is to use the BCELoss loss function to measure the difference between the probability and the true label. We select Adam as the optimizer of G and D and set the training iteration period to 500 and the learning rate to 0.0001.

In summary, eleven models are obtained in this study: ELM-sig, ELM-sin, ELM-hardlim, BPNN-tansig, BPNN-logsig are compiled in MATLAB, and RNN, Transformer, CNN, GAN ELM, BPNN are compiled in Python.

Results

Spectral analysis. FT-NIR spectroscopy is generated by the transition of molecular vibration from the ground state to the higher energy level in overtones and combination modes due to the non-resonance of molecular vibration. It mainly records the overtones and combination modes absorption of the vibration of hydrogen-containing groups, covering the composition and molecular structure information of most types of organic compounds[48]. Due to the difference in the absorption wavelength of organic compounds, FT-NIR spectroscopy is suitable for the determination and analysis of hydrocarbon organic compounds.

Figure 2 is the FT-NIR spectrum of the fennel after baseline correction. (a) is the sample of fennel from Yumen, Gansu, and (b) is the sample of fennel from Turpan, Xinjiang. In Fig. 2, it can be observed that there are peaks at 4260, 4330, 5155, 5670, 5808, 6840 cm^{-1}, etc., and the peak shape is relatively sharp, indicating that the information contained in this place is relatively rich. The region between 5000 and 4000 cm^{-1} is mainly the frequency absorption of C–H, and this region is generally considered to be the characteristic absorption band of sugar, protein and starch[49–51]. In the vicinity of 5155 cm^{-1}, there are two vibrations due to O–H stretching and O–H deformation, which can be speculated as water absorption zone[52], which may be because the dried fennel still contains a small amount of water. The FT-NIR spectrum at 7500 and 5500 cm^{-1} is mainly the stretching vibration of C–H in CH_3– and CH_2– groups. It can be seen from Figure 1 that the peak distribution and trend of the FT-NIR spectra of the samples of Yumen fennel in Gansu and Turpan fennel in Xinjiang are similar, indicating that the sample composition and related information are similar. Spectral peaks of individual bands

Wavenumber (cm^{-1})	Assignment	Substances related including in fennel
4260	O–H, N–H and C–O bands	Essential oils
4330	C–H of Lipids	Hydroxy fatty acids
4830	Amide	Proteins
5155	C–H, N–H, O–H of water molecule	Water
5670	$–CH_2$	Steroids
5808	$–CH_3$ of polyphenols	Polyphenols
6840	polyamides	Flavonoids

Table 1. Waveband and composition distribution of major NIR peaks of the fennel.

Principal component	Contribution rate(%)	Cumulative contribution rate(%)
1	85.93	85.93
2	6.22	92.14
3	4.80	96.95
4	1.45	98.40
5	0.51	98.91
6	0.35	99.26

Table 2. The contribution rates of the first six features in PCA.

(such as 4330 and 5155 cm^{-1}, etc.) are different, indicating that there are significant differences in the percentage content of various oils and substances such as water content contained in fennel from different origins, which is also consistent with the conclusions reached in some previous studies[53,16]. Therefore, it is feasible to use FT-NIR spectroscopy to classify and identify fennel in Gansu and Xinjiang. As can be seen from Table 1, the main substances in fennel include nutrients such as proteins and fatty acids, polyphenols (hydroxyl and –CH groups are mostly phenols) other antioxidant substances and flavonoids, etc.[12,15,54].

PCA feature extraction. If the number of principal components is too large, it is easy to introduce noise and redundant data[55]. In this experiment, we selected the first 6-dimensional data of the fennel spectral data as the principal components, and the cumulative variance contribution rate has reached 99.26%. The specific information of the contribution rate is shown in Table 2. Their first three PCs score are shown in Fig. 3, and the drawing software used is Origin Pro 2019b.

Model evaluation. In this study, the Gansu fennel sample is regarded as a positive sample, and the Xinjiang fennel sample is regarded as a negative sample. After eleven models of ELM-sig, ELM-sin, ELM-hardlim, BPNN-tansig, and BPNN-logsig, RNN, Transformer, GAN, BPNN, CNN, ELM are obtained, each model running 30 times, record the specificity, sensitivity, accuracy, and model of every run time, 30 times experiment after calculating the average of the index record in Table 3.

It can be seen from Table 3 that ELM- and BPNN-based algorithms have similar results for evaluation indexes in different programming languages, and even the running time of Python is slightly faster than that of MATLAB. The accuracy reflects the proportion of the number of correctly classified samples to the total number of samples. The accuracy of the BPNN model is more than 96%, and the accuracy of the ELM, RNN, GAN and Transformer models are above 98%, especially the ELM model with Hardlim as the activation function has the best classification effect with 100% accuracy and the fastest running time. In the ELM model, the selected activation function is different, and the running time is also different, but the difference is not obvious. Among them, the ELM model with Sine as the activation function has the shortest running time, indicating that the model has the fastest classification speed. It can be clearly concluded from the results in Table 3 that the running time of the ELM model is shorter than that of the BPNN and CNN models, and the classification accuracy of the ELM model is higher. In addition, the ELM model is faster in achieving the same high classification accuracy as RNN, GAN and Transformer models.

In order to further verify the reliability of the model's classification of fennel origin, we introduced a receiver operating characteristic curve (ROC) curve to evaluate it. The horizontal and vertical coordinates of the ROC curve represent the specificity and sensitivity of the model, and the area under the curve (AUC) can be used as an index to evaluate the classification effect of the model. Where specificity indicates the proportion of correctly classified unqualified samples to the number of unqualified samples, and sensitivity is the proportion of correctly classified qualified samples to the number of qualified samples. The closer the AUC value is to 1, the better the effect of the classifier[56]. We used OriginPro 2019b to draw the ROC curve of each model in Fig. 4, and further obtained the average AUC value of each model. The results are recorded in Table 3. It can be seen that the AUC values of the ELM, RNN, Transformer, GAN and BPNN models are greater than 0.96, which indicates

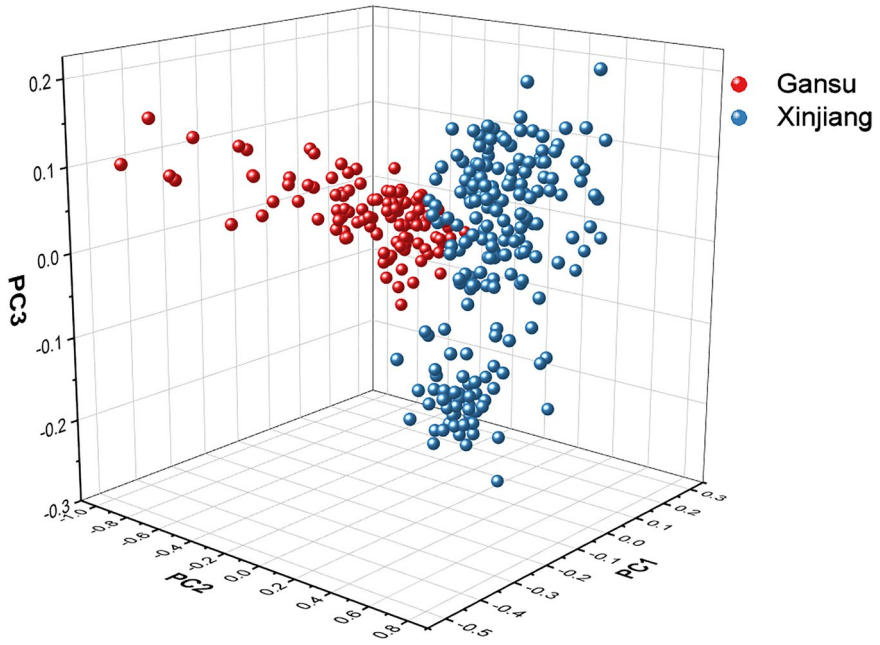

Figure 3. Three-dimensional scatter plot of three principal components of the fennel FT-NIR spectroscopy.

Languages	Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	Run time(s)	AUC
MATLAB	ELM-sig	97.89	99.18	98.73	0.05	0.98
	ELM-sin	97.63	99.51	98.84	0.04	0.98
	ELM-hardlim	**100.00**	**100.00**	**100.00**	**0.05**	**1.00**
	BPNN- tansig	95.92	97.47	96.91	0.51	0.96
	BPNN- logsig	98.07	97.17	97.43	0.44	0.97
Python	RNN	100.00	100.00	100.00	2.01	1.00
	Transformer	100.00	100.00	100.00	8.47	1.00
	GAN	100.00	100.00	100.00	126.22	1.00
	BPNN	98.75	93.33	97.89	0.45	0.96
	CNN	62.96	65.15	64.51	2.29	0.64
	ELM	**100.00**	**100.00**	**100.00**	**0.03**	**1.00**

Table 3. Experimental performance of each model comparison results. Significant values are in [bold].

their high reliability in classifying fennel origins in Yumen, Gansu and Turpan, Xinjiang. In Fig. 4, the AUC values of the ELM model are all greater than those of the BPNN model and CNN models, and the AUC value of the RNN, Transformer, GAN and ELM models with Hardlim as the activation function all reach 1.00, but the classification speed of the ELM model is the fastest, only 0.05s. In summary, the ELM model runs fast, has a high accuracy rate, a large AUC value, and the classification effect is better than the BPNN, RNN, Transformer, GAN and CNN models.

Conclusion

The kinds and contents of antioxidant substances, phenols, flavonoids and other substances in the fennel from the different producing areas are different[19]. In this study, the FT-NIR spectroscopy of the fennel combined with the machine learning algorithm is used for the first time to classify the fennel from different producing areas. The results show that the classification accuracy of the BPNN model is above 96%, the classification accuracy of the ELM, RNN, GAN and Transformer models is above 98%. The classification accuracy of the ELM model with the hardlim as the activation function can reach 100%. The classification speed of the ELM model is significantly faster than that of the RNN, GAN and Transformer models, with an average classification speed of 0.05s after 30 experiments. The experiments show that our proposed ELM model is more lightweight and faster in detection speed with guaranteed detection accuracy. Compared with the current mainstream deep learning models, the ELM model can combine the advantages of both high performance and low time cost, and at the same time solves the problem of poor detection accuracy and large time cost spent in the detection of massive data, which is more valuable for large-scale classification tasks in practical applications. The results of this study

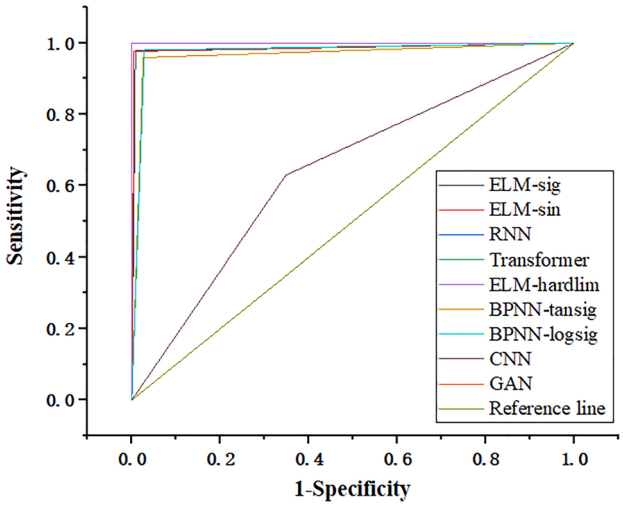

Figure 4. AUC curves of each model.

show that the use of the FT-NIR spectroscopy with simple sample preparation and fast detection speed, combined with machine learning algorithms, can achieve rapid identification of the fennel from different origins, thereby helping consumers better identify high-quality products and preventing unscrupulous merchants from shoddy behavior helps the market supervise related industries. In addition, this research technology should be introduced to other food classifications of different origins, providing new ideas for the intelligent supervision of the origin of various foods in the future.

Data availability

All data generated or analysed during this study are included in this published article [and its supplementary information]. S1 File. In this file, 200 samples were collected from Lianyungang Kaihao Tong Trading Co. in Xinjiang, and 116 samples were collected from Gansu Yumen Xiaosannong Taobao online store.

Received: 17 March 2022; Accepted: 1 August 2022
Published online: 10 August 2022

References

1. Thippeswamy, N. B. & Naidu, K. A. Antioxidant potency of cumin varieties-cumin, black cumin and bitter cumin-on antioxidant systems. *Eur. Food Res. Technol.* **220**, 472–476 (2005).
2. Hosseini, S., Ramezan, Y. & Arab, S. A comparative study on physicochemical characteristics and antioxidant activity of sumac (rhus coriaria l.), cumin (cuminum cyminum), and caraway (carum carvil) oils. *J. Food Meas. Charact.* **14**, 3175–3183 (2020).
3. Dinparvar, S. *et al.* A nanotechnology-based new approach in the treatment of breast cancer: Biosynthesized silver nanoparticles using cuminum cyminum l. seed extract. *J. Photochem. Photobiol. B* **208**, 111902. https://doi.org/10.1016/j.jphotobiol.2020.111902 (2020).
4. Riasat, M., Heidari, B., Pakniyat, H. & Jafari, A. A. Assessment of variability in secondary metabolites and expected response to genotype selection in fenugreek (Trigonella spp.). *Ind. Crops Prod.* **123**, 221–231. https://doi.org/10.1016/j.indcrop.2018.06.068 (2018).
5. Archangi, A., Heidari, B. & Mohammadi-Nejad, G. Association between seed yield-related traits and cdna-aflp markers in cumin (cuminum cyminum) under drought and irrigation regimes. *Ind. Crops Prod.* **133**, 276–283 (2019).
6. Milan, K., Dholakia, H., Tiku, P. K. & Vishveshwaraiah, P. Enhancement of digestive enzymatic activity by cumin (cuminum cyminum l.) and role of spent cumin as a bionutrient—sciencedirect. *Food Chem.* **110**, 678–683 (2008).
7. Olgun, O. & Yildiz, A. O. Effect of dietary supplementation of essential oils mixture on performance, eggshell quality, hatchability, and mineral excretion in quail breeders. *Environ. Sci. Pollut. Res.* **21**, 13434–13439 (2014).
8. Izabela, K. & Wei, Z. Anthocyanins-more than nature's colours. *J. Biomed. Biotechnol.* **2004**, 239–240 (2004).
9. Bagirova, M. *et al.* Investigation of antileshmanial activities of cuminum cyminum based green silver nanoparticles on l. tropica promastigotes and amastigotes in vitro. *Acta Trop.* **208**, 105498. https://doi.org/10.1016/j.actatropica.2020.105498 (2020).
10. Pereira, A. S. P., Banegas-Luna, A. J., Pea-García, J., Pérez-Sánchez, H. & Apostolides, Z. Evaluation of the anti-diabetic activity of some common herbs and spices: Providing new insights with inverse virtual screening. *Molecules* **24**, 4030 (2019).
11. Diaz-Maroto, M. C., Perez-Coello, M. S., Esteban, J. & Sanz, J. Comparison of the volatile composition of wild fennel samples (foeniculum vulgare mill.) from central Spain. *J. Agric. Food Chem.* **54**, 6814–6818 (2006).
12. Bettaieb, I. *et al.* Essential oils and fatty acids composition of Tunisian and Indian cumin (cuminum cyminum l.) seeds: A comparative study. *J. Sci. Food Agric.* **91**, 2100–2107 (2011).
13. Toma, C. C., Pancan, I. B., ChiriŢă, M., Vata, F. M. & Zamfir, A. D. Electrospray ionization tandem mass spectrometric investigation of essential oils from Melissa officinalis (Labiatae Family) and Pellargonium ssp. (Geraniaceae Family). In *Applications of Mass Spectrometry in Life Safety* (eds Popescu, C. *et al.*) 213–220 (Springer, Netherlands, Dordrecht, 2008).
14. Özbek, H. *et al.* Hepatoprotective effect of foeniculum vulgare essential oil. *Fitoterapia* **74**, 317–319 (2003).
15. Bettaieb Rebey, I. *et al.* Comparative assessment of phytochemical profiles and antioxidant properties of Tunisian and Egyptian anise (pimpinella anisum l.) seeds. *Plant Biosyst.* **11263504**(2017), 1403394 (2017).
16. Yaldiz, G. & Camlica, M. Variation in the fruit phytochemical and mineral composition, and phenolic content and antioxidant activity of the fruit extracts of different fennel (foeniculum vulgare l.) genotypes. *Ind. Crops Prod.* **142**, 111852 (2019).

17. Merah, O. *et al.* Biochemical composition of cumin seeds, and biorefining study. *Biomolecules*https://doi.org/10.3390/biom10071054 *(2020).*
18. El Sheikha, A. F. How to determine the geographical origin of food by molecular techniques. In *Molecular Techniques in Food Biology*. https://doi.org/10.1002/9781119374633.ch1 (John Wiley & Sons, Ltd, 2018).
19. Kelly, S., Heaton, K. & Hoogewerff, J. Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. *Trends Food Sci. Technol.* **16**, 555–567 (2005).
20. Porep, J. U., Kammerer, D. R. & Carle, R. On-line application of near infrared (nir) spectroscopy in food production. *Trends Food Sci. Technol.* **46**, 211–230 (2015).
21. Diniz, P., Gomes, A. A., Pistonesi, M. F., Band, B. & Araújo, M. Simultaneous classification of teas according to their varieties and geographical origins by using nir spectroscopy and spa-lda. *Food Anal. Methods* **7**, 1712–1718 (2014).
22. Vitale, R. *et al.* A rapid and non-invasive method for authenticating the origin of pistachio samples by nir spectroscopy and chemometrics. *Chemom. Intell. Lab. Syst.* **121**, 90–99 (2013).
23. Laroussi-Mezghani, S. *et al.* Authentication of Tunisian virgin olive oils by chemometric analysis of fatty acid compositions and nir spectra. comparison with Maghrebian and French virgin olive oils. *Food Chem.* **173**, 122–132 (2015).
24. Pei, Y. F., Zuo, Z. T., Zhang, Q. Z. & Wang, Y. Z. Data fusion of fourier transform mid-infrared (mir) and near-infrared (nir) spectroscopies to identify geographical origin of wild paris polyphylla var. yunnanensis. *Molecules* **24**, 2559 (2019).
25. Anyidoho, E. K., Teye, E. & Agbemafle, R. Nondestructive authentication of regional and geographical origin of cocoa beans by using handheld nir spectroscopy and multivariate algorithm. *Anal. Methods* **12**, 4150–4158 (2020).
26. Zhao, H., Guo, B., Wei, Y. & Bo, Z. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. *Food Chem.* **138**, 1902–1907 (2013).
27. Liang, X. Y., Li, X. Y. & Wu, W. J. Classification of floral origins of honey by nir and chemometrics. *Adv. Mater. Res.* **605–607**, 905–909 (2013).
28. Yang, B., Chen, C., Chen, F., Chen, C. & Lv, X. Identification of cumin and fennel from different regions based on generative adversarial networks and near infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **260**, 119956 (2021).
29. Yan, T., Duan, L., Chen, X., Gao, P. & Xu, W. Application and interpretation of deep learning methods for the geographical origin identification of radix glycyrrhizae using hyperspectral imaging. *RSC Adv.* **10**, 41936–41945 (2020).
30. Liu, W., Liu, C., Yu, J., Zhang, Y. & Zheng, L. Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics. *Food Chem.* **251**, 86–92 (2018).
31. Huang, G.-B., Ding, X. & Zhou, H. Optimization method based extreme learning machine for classification. *Neurocomputing* **74**, 155–163. https://doi.org/10.1016/j.neucom.2010.02.019 (2010).
32. Kutsanedzie, F., Chen, Q., Hao, S. & Wu, C. In situ cocoa beans quality grading by near-infrared-chemodyes systems. *Anal. Methods* **9**, 5455–5463 (2017).
33. Zheng, W., Fu, X. & Ying, Y. Spectroscopy-based food classification with extreme learning machine. *Chemom. Intell. Lab. Syst.* **139**, 42–47 (2014).
34. Serag, A., Baky, M. H., Döll, S. & Farag, M. A. UHPLC-MS metabolome based classification of umbelliferous fruit taxa: A prospect for phyto-equivalency of its different accessions and in response to roasting. *RSC Adv.* **10**, 76–85. https://doi.org/10.1039/C9RA07841J (2020).
35. Yan, Z. *et al.* Rapid identification of benign and malignant pancreatic tumors using serum Raman spectroscopy combined with classification algorithms. *Optik* **208**, 164473. https://doi.org/10.1016/j.ijleo.2020.164473 (2020).
36. Khan, S. *et al.* Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. *Photodiagn. Photodyn. Ther.* **23**, 89–93. https://doi.org/10.1016/j.pdpdt.2018.05.010 (2018).
37. Devos, O., Downey, G. & Duponchel, L. Simultaneous data pre-processing and svm classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food Chem.* **148**, 124–130 (2014).
38. Richter, B., Rurik, M., Gurk, S., Kohlbacher, O. & Fischer, M. Food monitoring: Screening of the geographical origin of white asparagus using ft-nir and machine learning. *Food Control* **104**, 318–325 (2019).
39. Wenjing, L., Zhaotian, S., Jinyu, C. & Chuanbo, J. Raman spectroscopy in colorectal cancer diagnostics: Comparison of pca-lda and pls-da models. *J. Spectrosc.* **2016**, 1–6 (2016).
40. Bartlett, P. L. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Trans. Inf. Theory* (1998).
41. Liu, D. *et al.* Lychee variety discrimination by hyperspectral imaging coupled with multivariate classification. *Food Anal. Methods* **7**, 1848–1857 (2014).
42. Suliman, A. & Yun, Z. A review on back-propagation neural networks in the application of remote sensing image classification. *J. Earth Sci. Eng.* **19**, 52–65 (2015).
43. Magwaza, L. S. *et al.* Nir spectroscopy applications for internal and external quality analysis of citrus fruit-a review. *Food Bioprocess Technol.* **5**, 425–444 (2012).
44. Zhou, Y., Wang, Y. & Yao, Q. Segmentation of rice disease spots based on improved bpnn. In *2010 International Conference on Image Analysis and Signal Processing* 575–578 (IEEE, 2010).
45. Hoogi, A., Mishra, A., Gimenez, F., Dong, J. & Rubin, D. L. Natural language generation model for mammography reports simulation. *IEEE J. Biomed. Health Inf.*https://doi.org/10.1109/JBHI.2020.2980118 *(2020).*
46. Wang, Z., Peng, D., Shang, Y. & Gao, J. Autistic spectrum disorder detection and structural biomarker identification using self-attention model and individual-level morphological covariance brain networks. *Front. Neurosci.* 1268 (2021).
47. Rong, D., Wang, H., Ying, Y., Zhang, Z. & Zhang, Y. Peach variety detection using vis-nir spectroscopy and deep learning. *Comput. Electron. Agric.* **175**, 105553 (2020).
48. Cen, H. & He, Y. Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends Food Sci. Technol.* **18**, 72–83. https://doi.org/10.1016/j.tifs.2006.09.003 (2007).
49. Gierlinger, N., Schwanninger, M. & Wimmer, R. Characteristics and classification of Fourier-transform near infrared spectra of the heartwood of different larch species (larix sp.). *J. Near Infrared Spectrosc.* **12**, 113 (2004).
50. Iñón, F. A., Llario, R., Garrigues, S. & de la Guardia, M. Development of a pls based method for determination of the quality of beers by use of nir: Spectral ranges and sample-introduction considerations. *Anal. Bioanal. Chem.* **382**, 1549–1561 (2005).
51. León, L., Kelly, J. D. & Downey, G. Detection of apple juice adulteration using near-infrared transflectance spectroscopy. *Appl. Spectrosc.* **59**, 593 (2005).
52. Chen, Q., Zhao, J. & Hao, L. Study on discrimination of roast green tea (camellia sinensis l.) according to geographical origin by ft-nir spectroscopy and supervised pattern recognition. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **72**, 845–850 (2009).
53. Merah, O. *et al.* Biochemical composition of cumin seeds, and biorefining study. *Biomolecules* **10**, 1054 (2020).
54. Koohsari, S., Sheikholeslami, M. A., Parvardeh, S., Ghafghazi, S. & Amiri, S. Antinociceptive and antineuropathic effects of cuminaldehyde, the major constituent of cuminum cyminum seeds: Possible mechanisms of action. *J. Ethnopharmacol.* **255**, 112786 (2020).
55. Martin, F. L. *et al.* Identifying variables responsible for clustering in discriminant analysis of data from infrared microspectroscopy of a biological sample. *J. Comput. Biol.* **14**, 1176–1184 (2007).
56. Faraggi, D. & Reiser, B. Estimation of the area under the roc curve. *Stat. Med.* **21**, 3093 (2002).

Acknowledgements
This work was supported by the National Key Research and Development Program of China (2019YFC1606100 and sub-program 2019YFC1606104), the Major science and technology projects of Xinjiang Uygur Autonomous Region (2020A03001 and sub-program 2020A03001-3), and the special scientific research project for young medical science (2019Q003).

Author contributions
E.Z., L.S., J.Y., C.C.,C.C and X.L. conceived and designed the study. E.Z., L.S., J.Y. and C.C. collected and pre-processed the data. All authors contributed to the manuscript. X.L.and C.C.reviewed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-17810-y.

Correspondence and requests for materials should be addressed to C.C. or C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.