Identification of Blueberry Producing Areas Based on CNN-SE and Near Infrared Spectroscopy

Guannan WANG[1], Shanshan TANG[1], Na WANG[2]*

1. Heilongjiang Bayi Agricultural University, Daqing 163319, China; 2. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China

Abstract [**Objectives**] This study was conducted to realize the rapid and nondestructive identification of blueberry producing areas and protect benefits of high-quality blueberry brands. [**Methods**] Five types of blueberries from different regions were selected as experimental subjects, and spectral analysis techniques were combined with deep learning. Firstly, standard normal variable transform (SNV) and convolutional smoothing (SG) were used to deal with scattering noise and other issues in original spectral data. Secondly, due to a large amount of redundant information and high correlation between adjacent wavelengths in the collected spectra, continuous projection algorithm (SPA) and partial least squares regression (PLS) were combined for screening of features with RMSE as the indicator, and 40 feature variables were obtained. Finally, a convolutional network model CNN-SE integrating a Squeeze and Excitation (SE) attention mechanism module was constructed and compared with convolutional neural network (CNN), support vector machine (SVM), and BP neural network. [**Results**] The CNN-SE model had the best effect, with the accuracy and precision of the test set reaching 95% and 94.56%, respectively, and the recall and F_1 score reaching 93.94% and 94.24%, respectively. [**Conclusions**] The CNN-SE convolution network model can realize rapid, nondestructive and high-throughout identification of blueberry producing areas.

Key words Near infrared spectroscopy technology; Blueberry; Deep learning; Origin identification

DOI:10.19759/j.cnki.2164-4993.2025.01.013

As the "king of berries", blueberries have unique nutritional value, especially its components anthocyanins and ellagic acid, which give them a unique position in the field of healthy food[1-2]. In 2017, the International Food and Agriculture Organization (FAO) listed blueberries as one of the five healthy foods, and their unique nutritional value brought economic benefits. In order to promote the sustainable development of blueberry industry, it is crucial to establish and maintain regional brands. Excellent brands can not only enhance the added value of products, but also improve consumers' trust in quality[3]. However, with the deep development of economic globalization, trade barriers have been gradually broken, and goods can be circulated in all countries and regions, resulting in complicated market competition. Because most agricultural products such as blueberries cannot be effectively distinguished by appearance[4], safety and quality problems occur frequently, and the development of origin traceability technology has become very important.

At present, in the research on the identification methods of blueberry producing areas, the research on traditional methods has reached a high level, and can realize the identification of blueberry blueberry producing areas[5]. However, these methods are all based on chemistry, and show the advantage of high accuracy, but the disadvantage is that they cannot be widely popularized and applied and have weak practical significance. As an efficient modern spectral analysis technique, infrared spectroscopy has the characteristics of rapidity, no damage and high accuracy[6], and has been widely used in food, tobacco[7] and other fields in recent years. It can predict real targets by establishing a classification model between spectral data and target values, which directly determines the effectiveness of the final result of the experimental method, so it is of great significance to establish a high-performance model.

Modeling methods are divided into traditional machine learning and deep learning. Partial Least Squares Regression (PLS)[8] in machine learning algorithm is a linear modeling method, which is suitable for dealing with variables with clear linear relationship, but it has difficulties in expressing high-dimensional and complex spectral data well. If it is widely used in the field of near infrared spectroscopy, it will have certain limitations. Other machine learning nonlinear algorithms are only suitable for shallow nonlinear data, and the results are limited. Deep learning can learn complex and abstract information from high-dimensional data through multiple layers of nonlinear transformations. Especially since 2012, with the breakthrough of deep learning technology, more and more scholars have applied deep learning technology to the field of spectroscopy. Leng[9] proposed a new diagnosis technique which combined the multispectral technique and deep learning. The experiment showed that the combination of CNN-BiLSTM network improved the accuracy by 10%. Zhou[6] designed a framework of convolutional neural network (CNN-ATT) for one-dimensional data classification, and compared with common machine learning models, the results showed that the designed

Received: October 10, 2024 Accepted: December 20, 2024

Supported by Natural Science Foundation of Heilongjiang Province (LH2022E099); Daqing Guidance Fund for Science and Technology Planning Project (zd-2023-63); San Heng San Zong Support Program of Heilongjiang Bayi Agricultural University (ZRCPY202216).

Guannan WANG (1999-), male, P. R. China, master, devoted to research about near infrared spectroscopy.

* Corresponding author.

CNN-ATT deep learning model had the best performance. Zhou *et al.*[10] proposed a method to distinguish medicinal *Tsuga chinensis* (Franch.) E. Pritz. from different habitats by using near infrared spectroscopy and deep learning model, and the results showed that the accuracy of CNN in identifying *T. chinensis* from different habitats was 100%. From above scholars' research, it can be seen that the application of deep learning technology in the field of spectroscopy is feasible.

In this study, near infrared spectroscopy was applied to obtain the original spectral data of blueberries, which were then processed by spectral analysis techniques, and finally, a convolution neural network model (CNN-SE) integrating SE attention mechanism was constructed to realize nondestructive and high-throughout identification of blueberry producing areas.

Materials and Methods

Experimental instruments and sample collection

In this study, blueberry samples were collected from Dandong City, Daliangshan District, Chengjiang and Zhaotong City, and five different geographical regions in Chile, and 400 sets of data were accumulated. These data were divided into a training set (280 sets, accounting for 70%) and a test set (120 sets, accounting for 30%) for model training and verification. The near infrared spectrometer used was a portable diffuse reflection design, which was developed by Shenzhen Puyan Hulian Internet Technology Co., Ltd. before data collection, blueberries were cleaned with clear water and naturally dried to room temperature to ensure environmental consistency. In order to eliminate the potential deviation between equipment and experimental environment, whiteboard calibration procedure was implemented. Blueberry samples were placed in the center of the optical path of the spectrometer, so as to ensure non-destructive and comprehensive reception of spectral irradiation in the 900 – 1 700 nm band, with the interval accurate to 1.73 nm, and each spectrum recorded 355 data points, in order to capture the spectral characteristics of blueberries in detail.

Spectral data preprocessing method

In the collection and analysis of blueberry spectral data, near infrared spectroscopy is often challenged by the fluctuation of experimental environment, the consistency of equipment, the difference of operation and the different physical characteristics of samples. These factors lead to noise and scattered light mixed in the original spectral data, which affects the accuracy of data and the construction efficiency of subsequent model. In order to improve data quality and ensure the prediction accuracy and generalization ability of the model, we took following two key steps to preprocess the data. Firstly, the convolution smoothing technique (SG) is applied to smooth the spectral curve and preserve the details of the peak shape to a certain extent through local polynomial fitting, so that the real absorption characteristics originally covered by noise can be clearly displayed. Secondly, targeting at baseline shift issues caused by sample size differences, uneven surface reflections or environmental fluctuations in spectral data, the standard normal variable transformation (SNV) was applied to perform standardization of data distribution, which eliminated systematic intensity variations and improved data consistency and analyzability, and thus provided a more reliable foundation for subsequent model building. These pretreatment steps are important links to improve the accuracy of near infrared spectroscopy analysis, which ensures the reliability of spectral data and the validity of the model[11], and then supports the accurate evaluation of blueberry quality.

Spectral wavelength extraction method

The continuous projection algorithm (SPA) was used to screen effective features[12]. SPA is an efficient and intuitive feature selection technique, which can significantly reduce the dimension of feature space while maintaining the distinguishing ability of features, which is very important to improving the efficiency and accuracy of the model, especially when dealing with high-dimensional data sets such as near-infrared spectral data, during which it can effectively reduce the computational burden and optimize the feature set.

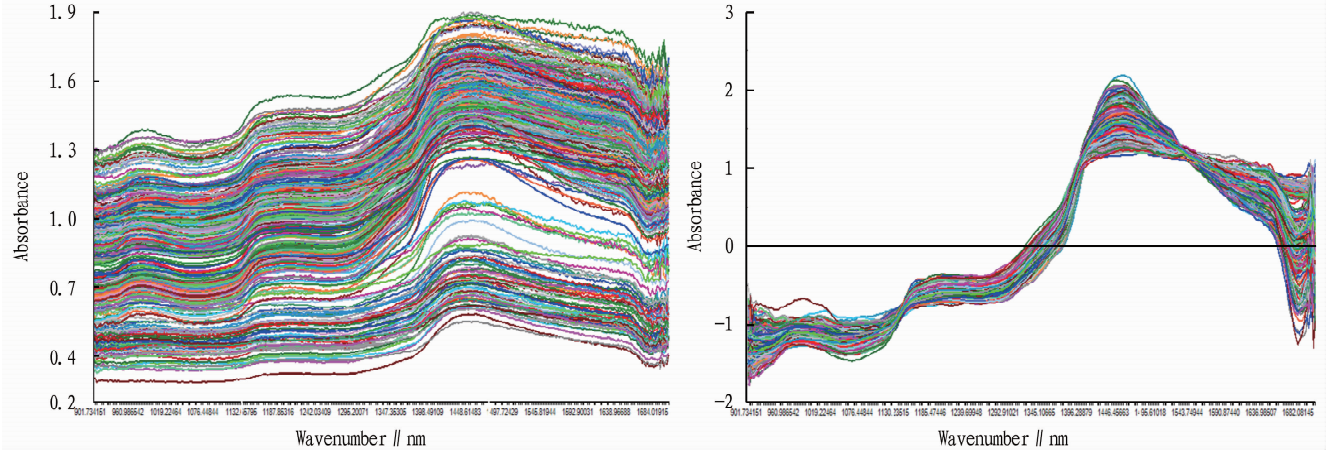
CNN-SE model establishment

In this study, a convolutional neural network for the classification of near-infrared spectral data sets was constructed and integrated into the SE attention mechanism module[13]. The attention mechanism can be called the resource allocation mechanism, and its idea draws lessons from the advantage that human beings give priority to the special parts of things when observing them. SE obtains channel-level statistical information through the global average pooling layer, and two fully connected layers generate channel weights, which reflect the importance of different channels and enable the model to dynamically adjust the weights of each channel in the feature map. It is helpful for the model to capture more abundant features and improve the ability of feature representation. Under the condition of limited computing power of neural network and high-latitude characteristics of near-infrared spectral data, more resource weights are allocated to some important features, so as to improve the accuracy of deep learning model with limited resources. Moreover, the computational overhead of SE is relatively small, and mainly involves global average pooling, two fully connected layers and a Sigmoid activation layer. These operations will not significantly increase the computational complexity of the model. Because the convolutional neural network gives the same weight to each part when extracting features, an attention mechanism module is introduced to make the convolutional network more focused on effective features and remove useless features, thereby effectively improving the performance of the model. Table 1 shows detailed parameters for building the model.

In order to comprehensively consider the model performance, four indicators including accuracy, precision, recall and F_1 score were used to evaluate model performance.

Table 1 CNN-SE model establishment

Network layer	Model parameter
Input layer	Near-infrared spectral data of blueberries
Convolution layer + activation layer	FilterSize = (3, 1) NumFilters = 32 Stride = (1, 1) activation function: relu_1
Convolution layer + activation layer	FilterSize = (3, 1) NumFilters = 64 Stride = (1, 1) activation function: relu_2
Global average pooling layer	Performing global average pooling on the output of the convolution layer, and reducing the features to vectors with fixed lengths.
Fully connected layer + activation layer	Realizing SE attention mechanism, and outputting 16 features via the first fully connected layer, activation function: relu_3
Fully connected layer + activation layer	Outputting 64 features via the second fully connected layer, activation function: sigmoid
Dot layer	Outputting SE attention mechanism, and adjusting the importance of various channels in features.
Seqfold Layer + flatten layer	Converting the features processed by SE attention mechanism into sequence format again.
Fully Connected Layer + output layer	Finally computing and classifying.



(a) Original near-infrared spectra; (b) Near-infrared spectra processed by SNV + SG.

Fig. 1 Near-infrared spectra of blueberries

Results and Analysis

Spectral analysis

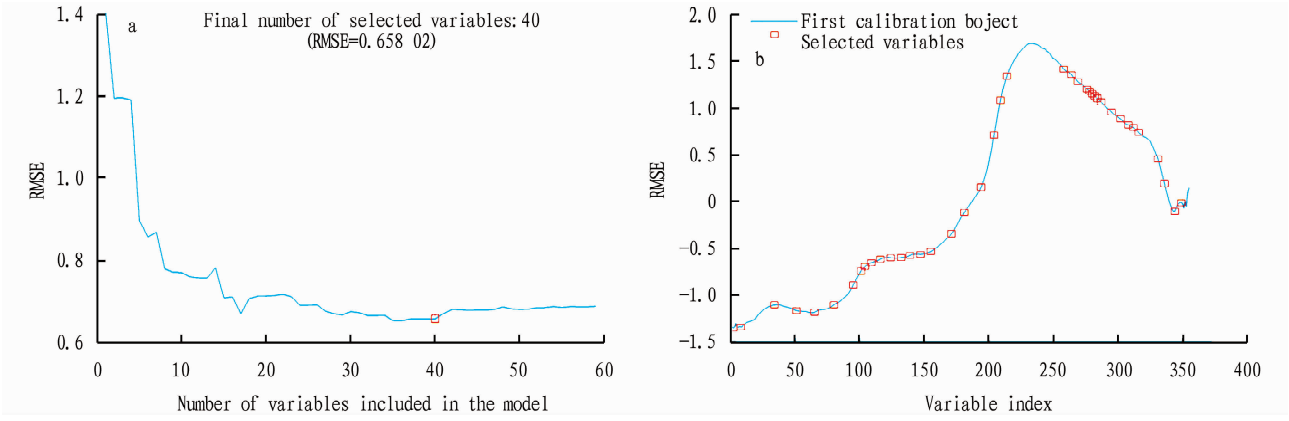
The original spectral data of 400 blueberry samples from different places shown in Fig. 1(a) show the consistency and difference of their spectral characteristics. The similar trend of these data showed that although the basic chemical composition of blueberries from different geographical locations was common, the difference in peak intensity implied a slight change in component content. In the 900 – 1 400 nm band, the rising trend of reflectivity might reflect the absorption characteristics of some components in blueberries with wavelength changes, while the downtrend in the 1 400 – 1 700 nm band might be related to the absorption characteristics of another group of chemical structures. Three significant absorption peaks were located at 980, 1 200 and 1 440 nm respectively, which were directly related to the vibration modes of specific chemical bonds in blueberries. The absorption peak near 980 nm was related to the second-order double-frequency vibration of O-H groups in water molecules, and the absorption peak near 1 200 nm was related to the stretching vibration of C-H bonds and their double frequency and combined frequency. The absorption peak around 1 440 nm was related to the first-order double-frequency stretching vibration of O-H, and the dense intersection of absorption peaks around 1 440 nm showed the complexity of this wavelength region, which might be due to slight changes of moisture and sugar components of blueberries from different origins, which increased the complexity of identifying the origin by spectral analysis. These differences not only reflected the natural variation among blueberry varieties, but also implied the effects of environment and growth conditions on the chemical composition of the fruit, which provides important spectral characteristics for in-depth study of blueberry quality and origin identification.

Analysis of spectral preprocessing and wavelength selection results

Spectral preprocessing The polynomial order for SG was set to 3, and a 5-point sliding window size was used. As shown in Fig. 1 (b), compared with the original spectra, the blueberry spectra subjected to convolution smoothing could effectively smooth the spectral curves and enhance the clarity of peaks, and the feature shapes of the original data were kept unchanged, which was very important for identifying key absorption features. SNV processing reduced systematic deviation caused by sample size difference and optical path change through standardization, making the spectra of different samples more consistent, which was convenient for subsequent analysis. In this study, both SG and SNV were adopted to process the original near-infrared spectral data of blueberries, and the spectral shape after SG + SNV processing inherited the

advantages of the two algorithms.

Analysis of wavelength selection results In the aspect of wavelength selection, SPA, as an unsupervised feature selection method, can initially screen representative bands, but it is not sufficient to construct efficient prediction models on its own. Combining SPA with PLS, and using RMSE as an evaluation index, the wavelength selection can be optimized systematically. As can be seen from Fig. 2(a), with the number of wavelengths increasing, RMSE initially dropped significantly, indicating that initial feature selection had a great influence on the model. When the number of features increased to a certain number, the decreasing trend of RMSE slowed down, indicating that the information gain that extra wavelengths might bring was limited, and finally 40 features were determined.



(a) Feature screening quantity; (b) Distribution position of screening features.

Fig. 2 Feature screening diagram by SPA

Model result analysis

It could be seen from the model classification results in Table 2 that CNN-SE had the best effect. In specific, the accuracy and precision of the training set were 98.66% and 98.56%, respectively, and the recall and F_1 score were 98.54%. The accuracy and precision of the test set were 95% and 94.56%, respectively, and the recall and F_1 score were 93.94% and 94.24% respectively. Compared with CNN, the improvement of the training set was not obvious, but the improvement of the test set was higher. Specifically, the accuracy and precision were increased by 2.15% and 2.28%, respectively, and the recall and F_1 score were increased by 1.58% and 1.93% respectively. However, CNN was still better than the training results of the two machine learning methods, and the comprehensive effects achieved by the test sets of both SVM and BP did not reach 90%. In terms of SVM, the accuracy and precision of the training set were 89.66% and 89.08%, respectively, and the recall and F_1 score were 89.1% and 89.08%, respectively. The accuracy and precision of the test set were 80% and 80.06%, respectively, and the recall and F_1 score were 79.64% and 79.84%, respectively. BP performed relatively better, as the values of the training set were basically above 90%, but the accuracy of the test set and other indicators were still around 85%. The results of this study showed that the introduction of SE module really improved the training results of CNN model, and the adaptive weighting mechanism enhanced the ability of the model to learn complex patterns, making it better generalized to unknown data. After dynamically adjusting the feature weights, the dependence on specific features was reduced, and the risk of overfitting was reduced.

Table 2 Model classification result

No.	Classification model	Wavelength selection	Proportion//%	Sample division	Accuracy//%	Precision//%	Recall//%	F_1 score//%
1	CNN-SE	40	11	Training set	98.66	98.56	98.54	98.54
				Test set	95.00	94.56	93.94	94.24
2	CNN	40	11	Training set	98.33	98.10	98.02	98.05
				Test set	92.85	92.28	92.36	92.31
3	SVM	40	11	Training set	89.66	89.08	89.10	89.08
				Test set	80.00	80.06	79.64	79.84
4	BP	40	11	Training set	92.00	91.88	91.58	91.72
				Test set	85.00	84.90	84.36	84.62

Conclusions and Discussion

In this study, the original spectral data of blueberries were collected by a near infrared spectrometer. In order to eliminate the noise caused by the difference of environment, equipment and fruit, the original spectra were preprocessed by spectral analysis techniques. Meanwhile, 40 characteristic variables were screened by SPA algorithm, and finally, a CNN-SE convolution network model was established. Through comparison with other three

models, it was found that the model had the best effect, and could achieve fast, non-destructive and high-throughput identification of blueberry production areas.

References

[1] KALT W, CASSIDY A, HOWARD LR, *et al.* Recent research on the health benefits of blueberries and their anthocyanins[J]. Advances in Nutrition, 2020, 11: 224 -236.

[2] SILVA S, COSTA EM, VEIGA M, *et al.* Health promoting properties of blueberries: A review[J]. Critical reviews in food science and nutrition, 2020, 60: 181 -200.

[3] FRANCOIS G, FABRICE V, DIDIER M. Traceability of fruits and vegetables[J]. Phytochemistry, 2020, 173: 112291.

[4] LU SY. Study on the origin of cherries by Raman spectroscopy combined with pattern recognition[D]. Hangzhou: China Jiliang University, 2021. (in Chinese).

[5] KUANG L, NIE J, ZHANG J, *et al.* Discrimination of geographical origin of blueberries from three major producing areas of China using mineral element analyses[J]. Atomic Spectroscopy, 2021, 42: 91 -98.

[6] ZHOU L, ZHANG C, TAHA MF, *et al.* Wheat kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method[J]. Frontiers in Plant Science, 2020, 11: 575810.

[7] GENG Y, NI H, SHEN H, *et al.* Feasibility of an NIR spectral calibration transfer algorithm based on optimized feature variables to predict tobacco samples in different states[J]. Analytical Methods, 2023, 15: 719 -728.

[8] YUAN L, FU X, YANG X, *et al.* Non-destructive measurement of egg's haugh unit by Vis-NIR with iPLS-Lasso selection[J]. Foods, 2023, 12: 184.

[9] LENG H, CHEN C, CHEN C, *et al.* Raman spectroscopy and FTIR spectroscopy fusion technology combined with deep learning: A novel cancer prediction method[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 285: 121839.

[10] ZHOU D, YU Y, HU R, *et al.* Discrimination of Tetrastigma hemsleyanum according to geographical origin by near-infrared spectroscopy combined with a deep learning approach[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 23: 118380.

[11] MISHRA P, BIANCOLILLO A, ROGER JM, *et al.* Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques[J]. TrAC Trends in Analytical Chemistry, 2020, 132: 116045.

[12] LI J, ZHANG H, ZHAN B, *et al.* Nondestructive firmness measurement of the multiple cultivars of pears by Vis-NIR spectroscopy coupled with multivariate calibration analysis and MC-UVE-SPA method[J]. Infrared Physics & Technology, 2020, 104: 103154

[13] GUO HX, LI Y, CHEN LX, *et al.* Source-load extreme scenario identification method based on residual grouping convolutional neural network and multilevel attention mechanism[J]. Power System Technology, 2024, 0688: 1 -17. (in Chinese).

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU

(Continued from page 56)

Moreover, the microbial detection rate was low, and drugs were used empirically without etiological support sometimes. In view of the above phenomena, clinical pharmacists in the hospital have formulated a series of targeted and operable intervention measures. For example, clinicians are regularly organized to carry out training on the use of special-use antibiotics. With the help of information technology, a prereview system for inpatient medical orders and prescriptions has been loaded, and doctors can pop up a window to prompt the use of precautions, incompatibility, consultation and microbial inspection. Pharmacists can strengthen after-the-fact comments on hospitalized cases, and regularly feed back the comments to various departments. The cooperation among the information department, medical department, central laboratory and other departments is strengthened, and the use catalogue has been formulated to standardize the prescription authority of clinicians and strictly implement the consultation system of special-use antibiotics. Performance rewards and punishments and other measures are implemented to ensure the standardization and rationalization of the use of special-use antibiotics in the hospital and delay the occurrence of drug resistance.

References

[1] HU FP, GUO Y, ZHU DS, *et al.* Surveillance of drug resistance of bacteria in tertiary hospitals by CHINET in 2019[J]. Chinese Journal of Infection and Chemotherapy, 2020, 20(3): 234 -243. (in Chinese).

[2] HONG SN, ZHANG HQ, YAN XB, *et al.* Investigation on the use density of special-use antibiotics in hospitals in recent two years and its relationship with the drug resistance rate of common bacteria[J]. Chinese Journal of Clinical Rational Drug Use, 2024, 17(18): 158 -161. (in Chinese).

[3] LIU CL, CHEN D, XU HY, *et al.* Analysis of resistance of clinically common gram-negative bacilli to three common carbapenems[J]. Medical Laboratory Science and Clinices, 2019, 30(4): 5. (in Chinese).

[4] LU L, YANG TJ, CHEN GF. Analysis of clinical use of special-use antibiotics in a provincial hospital in Zhengzhou[J]. Journal of Henan Medical College, 2024, 36(2): 187 -193. (in Chinese).

[5] YANG S, DUAN QL, MEN X, *et al.* Rationality analysis of the use of special-use antibiotics in 274 inpatients in a hospital[J]. Anti-Infection Pharmacy, 2022, 19(9): 1276 -1279. (in Chinese).

[6] WANG X, HUANG L, YANG L, *et al.* Analysis on the application of special use antibiotics in a hospital from 2019 to 2021[J]. Strait Pharmaceutical Journal, 2024, 36(6): 74 -78. (in Chinese).

[7] HARRIOTT MM, NOVERR MC. Importance of *Candida*-bacterial polymicrobial biofilms in disease[J]. Trends Microbiol, 2011, 19(11): 557 -563.

[8] WANG WQ. Clinical significance of *Candida* detection in lower respiratory tract samples of patients with chronic obstructive pulmonary disease [D]. Shanghai: Shanghai Jiao Tong University, 2020. (in Chinese).

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU