ᔆ OPEN ACCESS

A Near-Infrared Spectroscopy Universal Model for Maize Moisture Based on Fused Wave-lengths

MingHao Dong[1], Siqi Liu[1], GuangYue Zhang[1], Xue Wang[1*], TieMin Ma[1]

[1]College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing, China

*Corresponding author: Xue Wang
College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing, China

| **Abstract** | **Review Article** |

Developing a stable and reliable universal model for different sample characteristics is a complex undertaking. During spectral calibration, variable shifts arising from sample characteristics and environmental conditions can compromise algorithmic accuracy. To address substantial errors in sample detection for universal models, a wavelength screening algorithm is incorporated into the pre-processing stage of model transfer. This removes wavelengths associated with irrelevant variables and information correction, retaining only effective variables as the final feature subset. This facilitates spectral data correction by model transfer algorithms, thereby realising universal modelling. Selecting CARS and RF, along with their combined wavelength screening methods, we constructed universal models for different maize traits in conjunction with the DS method. Experimental validation demonstrated that the combined wavelength screening methods outperformed single-wavelength screening models across all evaluation metrics in the universal models. Among 12 spectral datasets for distinct traits, the CARS-RF-DS model constructed by integrating features extracted from outlier sample sets—demonstrated optimal performance. Its correlation coefficient R^2 reached 0.9603, RMSEP was 0.0102, with an MAE of 0.0080. For the ear grain sample set, R^2 exceeded 0.93 and RPD surpassed 3.8, indicating that the CARS-RF-DS model possesses a degree of universality and generalisability.
Keywords: maize grain moisture, wavelength selection, feature fusion, quantitative analysis.

INTRODUCTION

In recent years, near-infrared spectroscopy has experienced rapid advancement, owing to significant improvements in instrument manufacturing technology alongside continuous progress in machine learning algorithms and data analysis software [1]. By employing appropriate spectral data processing methods, mathematical model relationships can be established between the near-infrared spectra of test samples and their physicochemical concentration or content values. This enables rapid, non-destructive, real-time monitoring of sample composition and content information, leading to its widespread application in food, pharmaceuticals, agricultural products, and other fields [2-6]. Typically, spectral models have a limited scope of applicability, being restricted to samples of the same origin as those used for model calibration. Should testing conditions or samples vary, established spectral models become unsuitable [7, 8]. Model transfer constitutes a technical approach employed in near-infrared quantitative analysis to maintain the universality of established models. Its fundamental purpose is to address spectral discrepancies arising from measurements taken on different instruments or under varying environmental conditions, as well as deviations in homogeneity between different samples [9]. Wavelength screening algorithms also play a crucial role in enhancing the predictive capability of universal models for near-infrared quantitative analysis [10]. Consequently, establishing universal models for near-infrared quantitative analysis by combining model transfer techniques with wavelength screening methods not only avoids resource wastage from model reconstruction but also facilitates the broader application of near-infrared spectral detection technology across other fields.

Addressing the limitation of NIR analytical models failing to accurately predict samples with different traits, CARS and RF screening methods were employed to obtain wavelength variables. Variable selection was conducted simultaneously on both master and slave spectra, with alternating eigenvalue alignment applied to chemically meaningful intervals in both spectra. Subsequently, a feature layer fusion strategy was applied to integrate the extracted feature bands. Combined with a direct correction algorithm, this

approach established a universal model for detecting moisture in maize grain with varying traits, thereby offering a novel methodology for moisture detection in maize breeding.

1 Relevant Theories
1.1 Random Frog (RF) Algorithm
RF draws upon the principles of reversible jump Markov Chain Monte Carlo (RJMCMC) algorithms. It achieves a pseudo-MCMC chain by implementing fixed-dimension and cross-dimensional movement searches within the model space. The probability of each variable being selected during iterations is calculated, with those exhibiting higher selection probabilities being prioritised as feature variables [11]. RF simulates the communication and cooperative behaviour exhibited by a population of frogs during foraging. It typically comprises two processes: local search and global exchange. Subpopulation local search alternates with global search across the entire population, thereby solving the problem.

1.2 Competitive Adaptive Re-weighted Sampling (CARS)
The CARS algorithm is a feature selection method for PLS model regression coefficients that combines Monte Carlo sampling with the principle of optimal model accuracy [12]. It emulates the principle of 'survival of the fittest' from Darwinian theory. At each iteration of the algorithm, variables with smaller weights are eliminated through adaptive weighted sampling, while those with larger weights are retained to form a new subset of variables. Multiple iterative calculations are then performed to select the wavelength variables within the subset that minimise the cross-validation root mean square error (Root Mean Square Error of Cross Validation, RMSECV) of the PLS model.

1.3 Direct Standardisation (DS)
The core objective of the DS algorithm is to achieve precise standardisation of the data to be calibrated while maintaining the stability and consistency of the calibration model [13]. This standardisation aims to significantly reduce discrepancies between the raw data and the standard data, ensuring data consistency and accuracy. This provides a more reliable foundation for subsequent spectral analysis and substance identification. Processing with the DS algorithm effectively minimises data variations, enhancing the comparability and usability of the spectral data.

2 CARS and RF-Based Fusion Wavelength Universal Model
2.1 Data Sampling
During data acquisition, near-infrared spectroscopy was performed on maize grains both on the ears and after threshing, yielding two spectral datasets: SZL and ZL respectively. Differences in sample characteristics can lead to issues such as baseline drift and background interference in the raw spectral data of the samples being analysed, affecting the precision of data analysis and the stability of modelling. To mitigate these issues, spectral preprocessing was performed using the SG convolution smoothing algorithm. Figure 1 shows the spectral data after SG smoothing preprocessing，(a) as shows spectral of the grains and (b) as shows spectral of grains on ears.
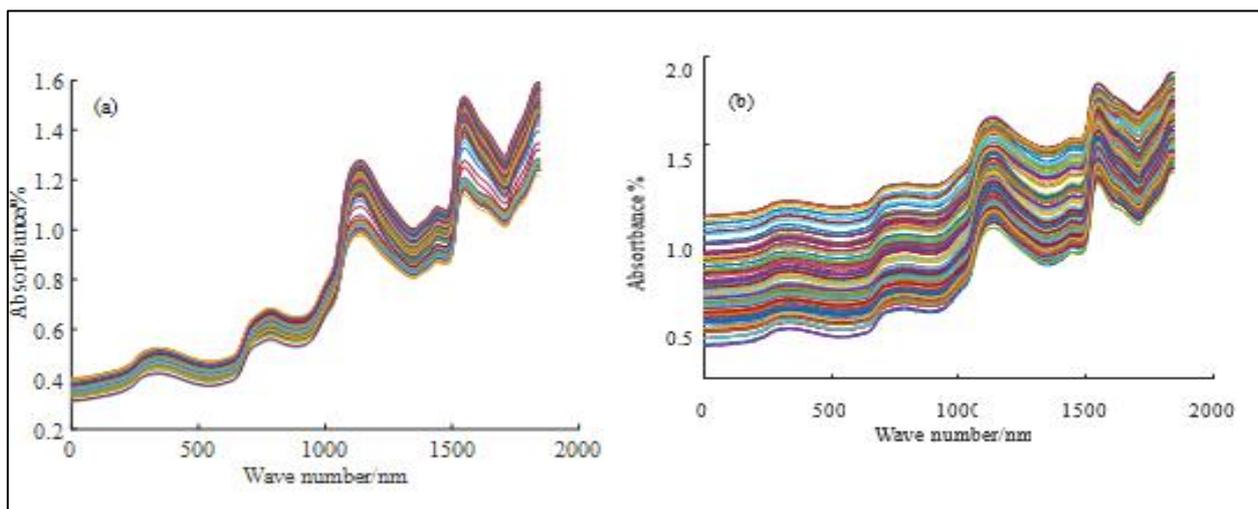

Fig. 1: Near-infrared spectroscopy of SG-pretreated grains and grains on ears

2.2 Wavelength Screening
In constructing generalised models, the abundance of spectral data features coupled with substantial interference information may yield suboptimal modelling outcomes. Consequently, wavelength screening methods are employed to extract characteristic wavelengths, thereby enhancing the predictive performance of generalised models. For the ZL and SZL spectral datasets representing distinct traits, two wavelength screening approaches—Competitive

Adaptive Re-Weighted Sampling (CARS) and Random Forest (RF)—were applied for feature extraction.

Competitive Adaptive Re-Weighted Sampling (CARS) is a feature variable selection method combining Monte Carlo sampling with PLS model regression coefficients. It selects points with greater weight through adaptive re-weighted sampling, choosing the wavelength variable subset corresponding to the minimum cross-validation root mean square error (RMSECV) as the characteristic wavelengths. The specific parameter settings for both datasets were: maximum principal component number 10, 50 Monte Carlo iterations, and an initial default feature count of 20. Given the random nature of CARS feature extraction, 10-fold cross-validation was employed to determine the optimal number of spectral bands. As shown in Figure 2(a), the RMSECV value for the ZL sample initially decreased gradually and stabilised. After achieving the minimum RMSECV value of 0.00069 at the 16th sampling iteration, further increases in sampling led to excessive removal of effective information bands, indicating that the optimal feature wavelength subset was selected at 16 samples. The selected feature wavelength distribution is shown in Figure 2(b). Observation of Figure 3(a) reveals that the SZL sample achieved its minimum RMSECV value of 0.00220 at iteration 18. Figure 3(b) presents the feature variable map obtained from the SZL sample following CARS feature wavelength extraction.
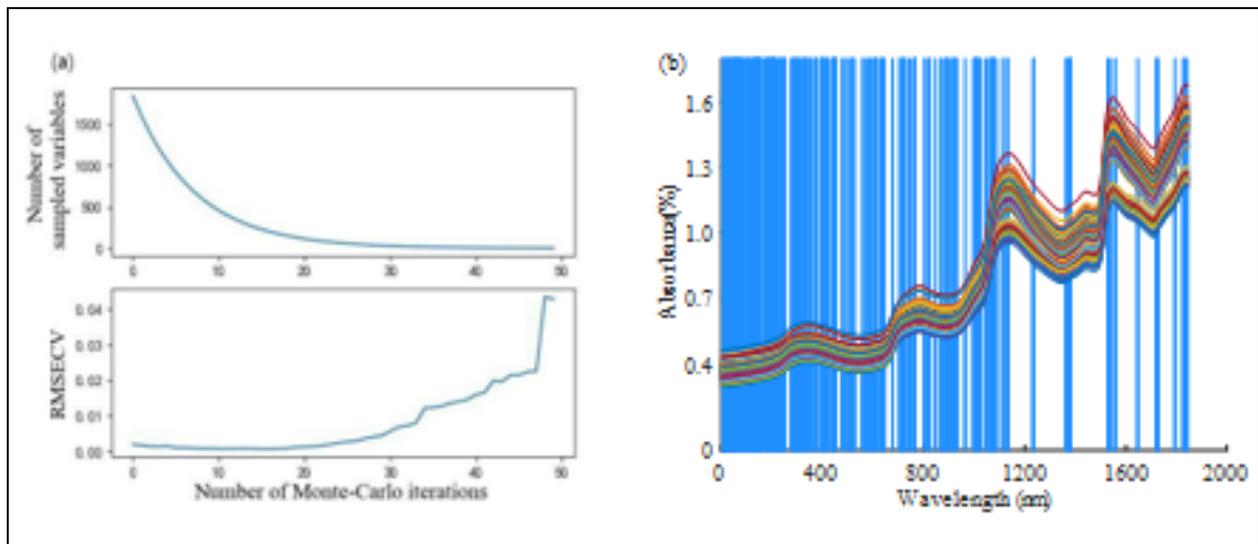

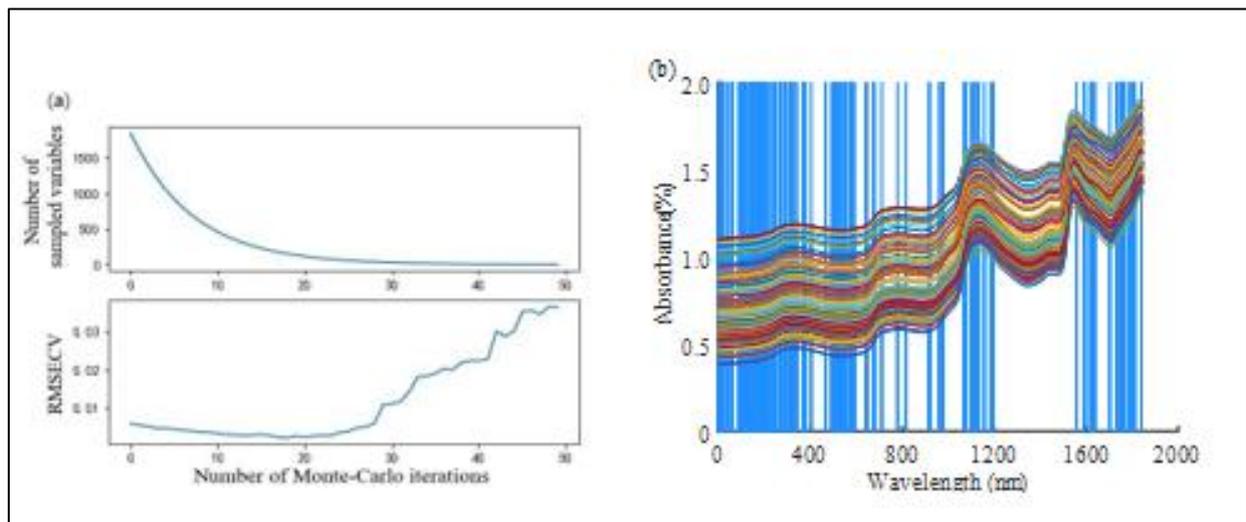**Fig. 2: ZL sample CARS band screening process (a) and Results (b)**

Fig. 3: SZL sample CARS band screening process (a) and Results (b)

As an ensemble algorithm, Random Forest enables feature selection based on computed feature importance. It calculates the mean squared error average using the out-of-bag (OOB) mean accuracy drop method as the feature importance metric. By setting a feature importance threshold, it filters feature variables to form a feature wavelength subset. The specific parameter settings are as follows: maximum number of latent

variables set to 20, 1000 iterations performed, data processing method set to centring, and feature variables selected where the selection probability exceeds the

threshold of 0.05. Figures 4(b) and (d) illustrate the feature variable plots obtained after RF feature wave length extraction.
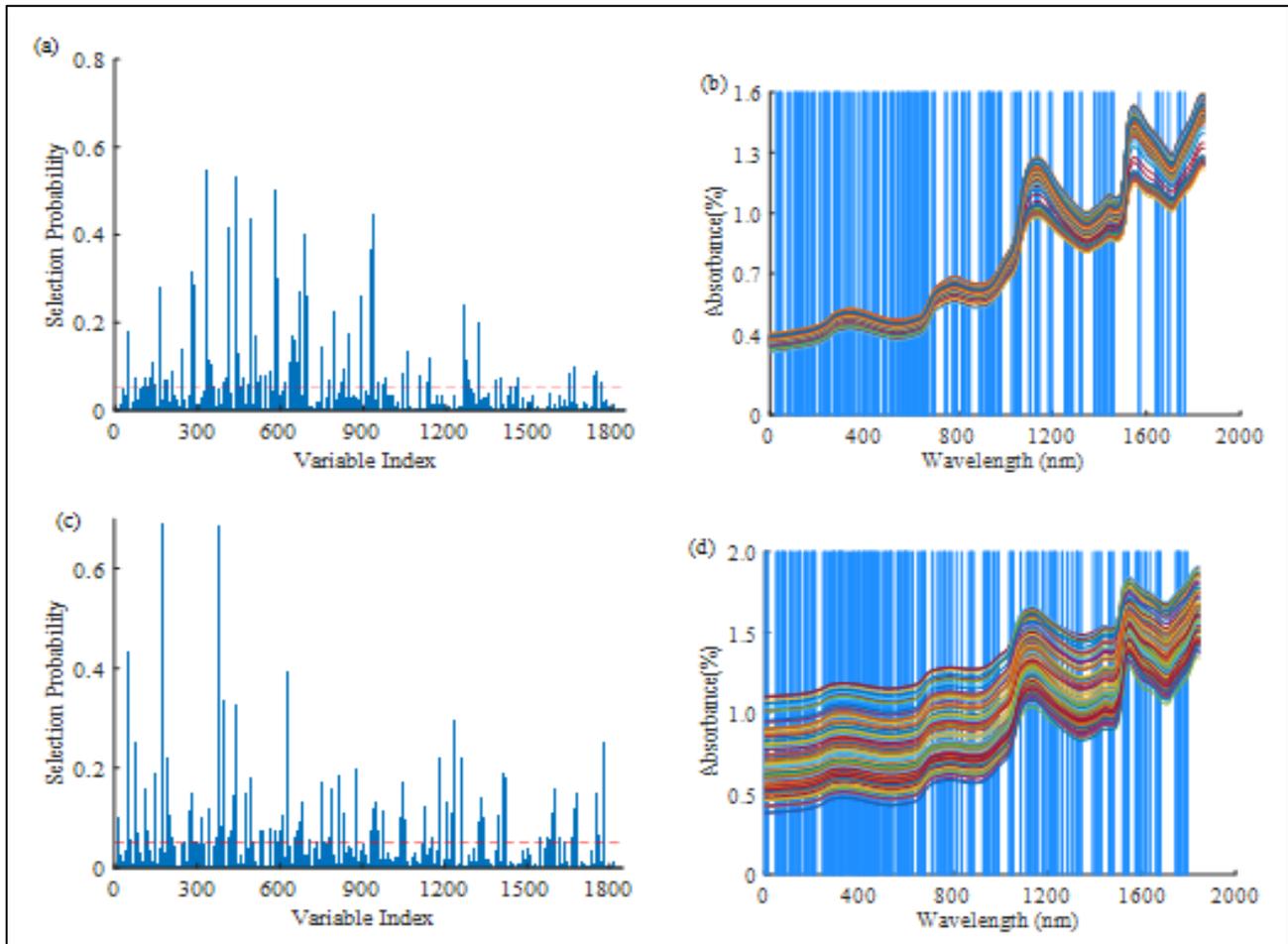

**Fig. 4: The wavelength feature extraction process and results based on RF**

2.3 Universal Model for Maize Grain Moisture Based on Fused Wavelengths

Model transfer constitutes a crucial approach for constructing universal models. During feature extraction to reduce spectral data dimensionality, the CARS and RF feature wavelength optimisation methods selected 199, 171, and 150, respectively, 208 wavelength points respectively. To mitigate potential errors arising from the two distinct trait datasets, post-feature extraction calibration and alignment of ZL and SZL sample spectra were performed. This involved selecting chemically meaningful intervals for both primary and secondary spectra, ensuring corresponding feature points align across different trait conditions and maintaining consistency in feature wavelength positioning. The merged wavelength feature set employed serial fusion for feature layer data integration. Three wavelength feature sets were combined with DS to establish PLS regression models, comparing and evaluating the performance of different wavelength selection methods

within a universal model. As shown in the table, model transfer via the DS algorithm significantly improved all metrics, with R^2 values consistently exceeding 0.9. The ZL sample's CARS-RF fusion model demonstrated markedly superior performance, achieving an R^2 of 0.9603, an RMSEP of merely 0.0102, and an MAE of just 0.0080, outperforming models built with other feature layer fusions. Compared to pre-fusion models, performance surpassed the best pre-fusion RF-PLS-DS model by 1.46% in correlation coefficient and exceeded the CARS-PLS-DS model by 4.75%, while also outperforming models built using single-wavelength selection. For the SZL sample set, the model demonstrated comparable performance. Compared to the best pre-fusion model (RF-PLS-DS), the R^2 value increased by nearly 1.4%, while RMSEP and MAE decreased by 24% and 18% respectively. The RPD improved by 18.7%, exceeding 3.8. This indicates that the universal model possesses practical applicability for both distinct sample types.

Table 1: Comparison of model transfer results of different wavelength screening methods and different dataset

样本集合	模型	预测集			
		R^2	RMSEP	MAE	RPD
	PLS	0.2109	0.1979	0.1917	0.2501
	RF-PLS	0.6870	0.0886	0.0693	0.5264
ZL	CARS-PLS	0.7657	0.1200	0.1162	0.4163
	CARS-DS-PLS	0.9128	0.0152	0.0128	3.2926
	RF-DS-PLS	0.9457	0.0125	0.0095	4.0036
	CARS-RF-DS-PLS	0.9603	0.0102	0.0080	4.9937
	PLS	0.2109	0.1979	0.1917	0.2501
	RF-PLS	0.6809	0.3265	0.3138	0.1345
SZL	CARS-PLS	0.6205	0.3871	0.3719	0.1320
	CARS-DS-PLS	0.8776	0.0175	0.0122	2.8168
	RF-DS-PLS	0.9212	0.0158	0.0105	3.2258
	CARS-RF-DS-PLS	0.9339	0.0120	0.0086	3.8292
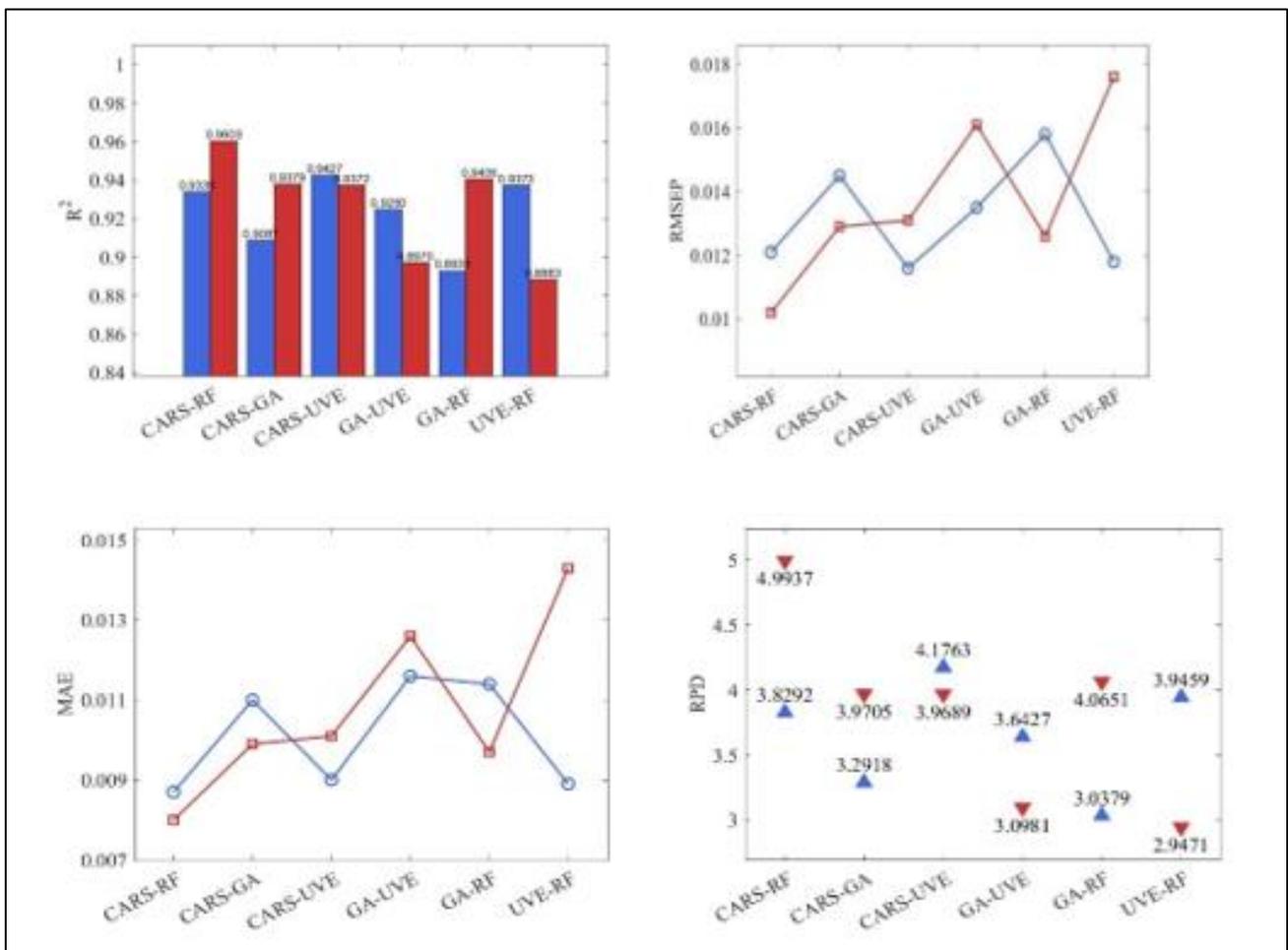
Fig 5: Comparison of model migration results after fusion of 12 sets of spectral data

2.4 Comparative Analysis of Model Performance

To further investigate the performance of the universal maize moisture model, the Genetic Algorithm (GA) and Uninformative Variable Elimination (UVE) variable selection methods were employed to construct universal models for comparative analysis. Feature information extracted via wavelength optimisation algorithms was combined in pairs, with serial fusion applied to integrate feature-layer data. This yielded 12 spectral datasets, each used to establish regression calibration models with DS and PLS respectively.

The model transfer results for the 12 combined spectral datasets are compared in Figures 5, where red indicates the feature-combined model performance for grain sample ZL, and blue denotes that for ear-attached grain SZL. Figure 5 clearly demonstrates that, compared to the universal model established using single-

wavelength selection, the feature layer fusion from the four wavelength selection methods better preserves the most representative spectral bands. Particularly, the CARS + RF fusion approach enhances the performance of the universal model. The constructed CARS-RF-DS universal model exhibits good universality, meeting the requirement for model universality across different traits.

3 CONCLUSIONS

(1)A universal model based on CARS + RF fusion wavelengths was established. The coefficient of determination for both on-ear grain moisture detection and post-threshing moisture detection exceeded 0.9, with an RPD exceeding 3.8. This meets the requirement for rapid grain moisture detection in maize breeding.

(2) To validate the effectiveness of the fusion wavelength model, feature information extracted by four wavelength screening algorithms—CARS, GA, RF, and UVE—was combined in pairs and compared with the CARS + RF fusion approach. This demonstrated that the constructed CARS-RF-DS universal model possesses favourable generalisability.

ACKNOWLEDGMENTS

REFERENCES

1. Peng Juan,SHAO Xueguang, CHU Ganghui.Rapid detection of Cymaiti apricot soluble solids content by miniature near infrared spectroscopy [J]. Chinese Journal of Analysis Laboratory,2023,42(10):1285-1291.
2. Wang Xue,Ma Tiemin,Yang Tao. Moisture quantitative analysis with small sample set of maize grain in filling stage based on near infrared spectroscopy [J]. Transactions of the Chinese Society of Agricultural Engineering, 2018, 34(13): 203-210.
3. GAO Feng, XING Ya-ge, LUO Hua-ping. Nondestructive Identification of Apricot Varieties Based on Visible/Near Infrared Spectroscopy and Chemometrics Methods [J]. Spectroscopy and Spectral Analysis, 2024, 44(1): 44-51.
4. SHAO Xiao-yu, JIANG Long-fa, ZHANG Ming. Application of Near-Infrared Spectroscopy in The Detection of Oil-Tea Camellia Seed Meal [J]. Infrared, 2023, 44(9): 38-45.
5. Li Jie, Wang Yuechao, Chen Pu. Application of Near-infrared Spectroscopy Analysis Technology in Chemical Industry [J]. Modern Chemical Research, 2024, (04): 96-99.
6. SUN Dai-qing, XIE Li-rong, ZHOU Yan. Application of SG-MSC-MC-UVE-PLS Algorithm in Whole Blood Hemoglobin Concentration Detection Based on Near Infrared Spectroscopy [J]. Spectroscopy and Spectral Analysis, 2021, 041(009): 2754-2758.
7. WANG Yun-hu, SUN Lei, WANG Cheng-zhong. Modeling Effect of Portable Near-Infrared Spectrometer in Identification of Fig Quality [J]. Journal of Shandong Polytechnic University, 2019, 033(004): 20-25.
8. MAO Xin-ran, XIA Jing-jing, XU Wei-xin. Study on Modeling Method of General Model for Measuring Three Quality Indexes of Pear by Handheld Near-Infrared Spectrometer [J]. Spectroscopy and Spectral Analysis, 2024, 44(02): 406-412.
9. Li M, Han D H, Lu D Q, *et al.,* Research Progress of Universal Model of Near-Infrared Spectroscopy in Agricultural Products and Foods Detection [J]. Spectorscopy and Spectral Analysis, 2022, 42(11): 3355-3360.
10. ZHANG Han, WANG Wen-zhe, HU Xiao-yan. Research on dynamic on-line monitoring method of moisture attribute in three honey-processed Chinese herbal slice based on in-situ general model [J]. Acta Pharmaceutica Sinica,2023,58(10):2890-2899.
11. Zhang Ziyue, Liu Xiao, Du Lili. Remote Sensing Inversion of Total Nitrogen in Hyperspectral Water Bodies Based on Random Frog Dimensionality Reduction and PSO-BPNN [J]. Laser & Optoelectronics Progress, 2025, 62(19).
12. TIAN Yongguo, LÜ Du, TANG Jianbo, *et al.,* Nondestructive Measurement of Passion Fruit Sugar Content Based on SiPLS and CARS Algorithms [J]. Food Research & Development, 2023, 44(20). 025.
13. LIU Zhijian, XIONG Zhixin, HU Yunchao.Research on near-infrared model transfer for lignin based on the combined direct standardization-piecewise direct standardization method [J]. Chinese Journal of Analysis Laboratory, 2023, 1-14.